A FUNCTIONAL EQUATION IN THE THEORY OF DYNAMIC PROGRAMMING AND ITS GENERALIZATIONS

Richard Bellman
Sherman Lehman

P-433

9 September 1953

The RAND Corporation
1700 MAIN ST. • SANTA MONICA • CALIFORNIA

Summary: ~~In this paper we study~~ Various analytic properties of ~~the~~ a particular functional equation are studied

~~$f(x,y) = \max[p_1(r_1x + r((1-r_1)x,y)),$~~

~~$p_2(r_2y + r(x,(1-r_2)y))]$~~,

together with a number of generalizations of discrete and continuous type.

TABLE OF CONTENTS

A FUNCTIONAL EQUATION IN THE THEORY OF DYNAMIC PROGRAMMING AND ITS GENERALIZATIONS

Richard Bellman and Sherman Lehman

§1. Introduction.

We propose in this paper to study a particular functional equation

$$f(x,y) = \text{Max}\ [\, p_1(r_1x + f((1-r_1)x,y)),\ p_2(r_2y + f(x,(1-r_2)y))\,], \qquad x,y \geq 0. \tag{1.1}$$

together with some of its generalizations and extensions.

The equation arises, as we shall show in the following section, in the following way: Let us assume that we possess two gold mines, Anaconda, which possesses an amount of gold in quantity x, and Bonanza, which possesses an amount y, together with one gold-mining machine. If the machine is used in the Anaconda mine, there is a probability p_1 that r_1x of the gold will be mined without damaging the machine, which means that the operation can be continued, and a probability $(1-p_1)$ that the machine will be damaged beyond repair and mine no gold. Similarly, the Bonanza mine has associated the probabilities q_1 and $(1-q_1)$ and the quantity r_2y. The problem is to determine the course of action which will maximize the expected amount of gold mined before the machine is damaged.

If we allow for a greater variety of outcomes, we obtain an extension of (1.1), namely,

$$f(x,y) = \max_{1\le i\le S}\left[\sum_{j=1}^{R} p_{ij}(r_{ij}x + s_{ij}y + f((1-r_{ij})x,(1-s_{ij})y)\right], \quad (1.2)$$

where

$$\begin{aligned} &(a)\quad x,y \ge 0 \\ &(b)\quad 0 \le r_{ij},\ s_{ij} \le 1 \\ &(c)\quad 0 < \sum_{j=1}^{R} p_{ij} < 1, \qquad i=1,2,\cdots,S. \end{aligned} \quad (1.3)$$

It is a simple step from this to the consideration of a continuum of outcomes, in which case the equation assumes the form

$$f(x,y) = \max_{1\le i\le S}\left[\int_0^{\infty}\left[r_i(t)x + s_i(t)y + f((1-r_i(t))x,\ (1-s_i(t))y)\right]dG_i(t)\right] \quad (1.4)$$

where

$$\begin{aligned} &(a)\quad x,y \ge 0 \\ &(b)\quad 0 \le r_i(t), s_i(t) \le 1 \\ &(c)\quad dG_i \ge 0,\ G_i(\infty) < 1, \qquad i=1,2,\cdots,S. \end{aligned} \quad (1.5)$$

The general problem will be one involving a number of different mines together with a number of machines of different performance. There is no difficulty in deriving similar, but more complicated equations,

$$f(p) = \max_{q} \left[\int_{R} [g(p,q,r) + f(T(p,q,r))]\, dG_q(r) \right] \tag{1.6}$$

where p is a point, $(x_1, x_2, \cdots x_n)$, and T(p,q,r) is a transformed point.

We shall begin our discussion by establishing an existence and uniqueness theorem which, while not nearly the most general which may be obtained, illustrates very clearly the methods that may be used. The basic method is, of course, that of successive approximations. We also discuss the dependence of f(p) upon parameters appearing in g and T.

As might be expected from the nonlinear nature of the functional equations, the solution of these equations is, in general, quite difficult to obtain or describe. Up to the present, only a handful have been completely resolved. In what follows, we shall consider (1.1), (1.2), and some immediate generalizations. The case where only a finite number of operations are permitted will also be treated.

Turning from the problem of maximizing the expected return, we shall consider the more general question of maximizing the expected value of some function of the return. The simplest analogue of (1.1) is then

$$\begin{aligned} f(x,y,t) = \max \big[& p_1 f((1-r_1)x, y, t + r_1 x) + (1-p_1)\phi(t), \\ & p_2 f(x, (1-r_2)y, t + r_2 y) + (1-p_2)\phi(t) \big] \end{aligned} \tag{1.7}$$

Under certain assumptions concerning $\phi(t)$, this equation can be solved, possessing a solution similar to that of (1.1). A particularly important case is that where $\phi(t) = e^{bt}$, $b > 0$. The

asymptotic form of $f(x,y,t)$ as $x,y \longrightarrow \infty$ can then be obtained.

It can be shown by means of counter-examples, cf. [8], that the difficulties encountered in the discrete formulation generating the preceding equations are due to the intricate form of the solution and that simple solutions, possessing an intuitive origin, are not to be obtained in all cases.

To overcome some of these difficulties sufficiently to obtain some approximate knowledge concerning the solutions, we have introduced continuous versions of the problems. These lead to problems in the calculus of variations which are fortunately sufficiently nonlinear to be susceptible to a variational attack. The problems are, however, not completely straightforward and require a non-classical type of argumentation.

Guided by our previous results, we consider in turn the two-choice, the three-choice, nonlinear utility and two-choice, finite-time problems, obtaining complete solutions.

We have treated only particular, simple cases of the continuous versions in order not to enmesh ourselves in conceptual difficulties. The general formulation requires a separate treatment which will be given elsewhere.

The central problem we have discussed in this paper is a particular maximization problem connected with multi-stage processes of deterministic and stochastic type. The general theory of these processes constitutes the theory of dynamic programming which has been discussed in a number of recent papers, [1] - [7].

§2. Mathematical Formulation.

To derive (1.1), let us set

$f(x,y)$ = expected amount of gold mined before the machine is damaged when A has an amount x, B has an amount y, and an optimal policy is pursued. (2.1)

If we choose to mine Anaconda, an operation we shall denote by A, with probability p_1 we obtain r_1x and the privilege of continuing; while with probability $(1-p_1)$ we obtain nothing. Since an optimal policy must have an optimal continuation, the expected return from an A-choice will be

$$E_A = p_1(r_1x + f((1-r_1)x,y)). \tag{2.2}$$

Similarly, the expected return from a B-choice is

$$E_B = p_2(r_2y + f(x,(1-r_2)y)). \tag{2.3}$$

Since our purpose is to maximize the expected return, we have

$$f(x,y) = \text{Max}\ (E_A,E_B), \tag{2.4}$$

which is precisely (1.1).

We can increase the possibilities without increasing the complexity of the equation. Let us assume that an A-choice has the following probabilities associated:

(a) p_1 = probability of obtaining r_1x and continuing

(b) p_2 = probability of obtaining 0 and continuing

(c) p_3 = probability of obtaining x and continuing (2.5)

(d) p_4 = probability of obtaining 0 and terminating the process

In a like manner, let B have the probabilities q_1, q_2, q_3, q_4 attached to its choice. Then we obtain

$$f(x,y) = \text{Max} \begin{bmatrix} A: & p_1[r_1x + f((1-r_1),xy)] + p_2 f(x,y) + p_3[x+f(0,y)] \\ B: & q_1[r_2y + f(x,(1-r_2)y)] + q_2 f(x,y) + q_3[y+f(x,0)] \end{bmatrix} \tag{2.6}$$

Since

$$f(x,0) = p_1[r_1x + f((1-r_1)x,0)] + p_2 f(x,0) + p_3x, \tag{2.7}$$

we have setting $f(x,0) = c_1x$,

$$f(x,0) = \frac{(p_1r_1 + p_3)x}{1-p_2-p_1(1-r_1)}, \tag{2.8}$$

and a corresponding expression for $f(0,y)$. The equation in (2.6) reduces to

$$f(x,y) = \text{Max} \begin{bmatrix} A: & c_{1A}x + c_{2A}y + p_1 f((1-r_1)x,y) \\ B: & c_{1B}x + c_{2B}y + q_1 f(x,(1-r_2)y) \end{bmatrix} \tag{2.9}$$

using (2.8) and the corresponding expression for $f(0,y)$ and solving for $f(x,y)$ where c_{1A}, c_{1B} are readily determined positive constants. The treatment of (2.9) is of the same order of difficulty as that of (1.1).

Let us now derive (1.7). Consider the same model as above in §1 and assume that we wish to maximize the expected value of $\phi(R)$ where ϕ is a given function and R is the total return obtained before the mining machine is damaged.

Setting

$$f(x,y,a) = \text{expected value of } \phi(R) \text{ obtained when A has x and B has y with an amount a already mined, using an optimal policy,} \tag{2.10}$$

we obtain, via the same argument as above, the functional equation of (1.7).

§3. Existence and Uniqueness.

Our first result is

Theorem 1. Consider the equation

$$f(p) = \max_{1 \le k \le m} \left[g_k(p) + h_k(p) f(T_k p) \right], \tag{3.1}$$

where we shall assume that

(a) The point p is restricted to a region R with the property that $p \in R$ implies that $T_k p \in R$.

(b) $|g_k(p)| \le c_1$ for $p \in R$ (3.2)

(c) $|h_k(p)| \le c_2 < 1$ for $p \in R$.

Under these conditions there is a unique bounded solution to (3.1).

Proof: Let $f_0(p)$ be an arbitrary bounded function for $p \in R$. Define

$$f_{n+1}(p) = \max_k \left[g_k(p) + h_k(p) f_n(T_k p) \right], \quad n=0,1,2,\cdots. \tag{3.3}$$

Let $k = k(n)$, dependent also upon p, be a value of k which furnishes the maximum, then

$$\begin{aligned} f_{n+1}(p) &= g_{k(n)}(p) + h_{k(n)}(p) f_n(T_{k(n)}(p)) \\ &\geq g_{k(n-1)}(p) + h_{k(n-1)}(p) f_n(T_{k(n-1)}(p)), \end{aligned} \tag{3.4}$$

and similarly

$$\begin{aligned} f_n(p) &= g_{k(n-1)}(p) + h_{k(n-1)}(p) f_{n-1}(T_{k(n-1)}(p)) \\ &\geq g_{k(n)}(p) + h_{k(n)}(p) f_{n-1}(T_{k(n)}(p)). \end{aligned} \tag{3.5}$$

From these relations we obtain for $n \geq 1$,

$$\begin{aligned} f_{n+1}(p) - f_n(p) &\geq h_{k(n-1)}(p) \left[f_n(T_{k(n-1)}(p)) - f_{n-1}(T_{k(n-1)}(p)) \right] \\ &\leq h_{k(n)}(p) \left[f_n(T_{k(n)}(p) - f_{n-1}(T_{k(n)}(p)) \right] \end{aligned} \tag{3.6}$$

Let us define

$$u_n = \sup_R |f_n(p) - f_{n-1}(p)|. \tag{3.7}$$

Using the bound given in (3.2c) we obtain from (3.6) the result

$$|f_{n+1}(p) - f_n(p)| \leq c_2 u_n, \tag{3.8}$$

whence $u_{n+1} \leq c_2 u_n$. This shows that the series $\sum_{n=1}^{\infty} u_n$ converges, which means that

$$\sum_{n=0}^{\infty} (f_{n+1}(p) - f_n(p)) \tag{3.9}$$

converges uniformly for $p \in R$. Hence $f_n(p)$ converges uniformly as $n \longrightarrow \infty$ to a function $f(p)$, a solution of the functional equation.

To establish the uniqueness of a bounded solution we proceed similarly. Let $F(p)$ be another solution of (1) and let k be an index which yields $f(p)$ and m be an index which yields F. Then, as above

$$f(p) = g_k(p) + h_k(p) f(T_k p) \geq g_m(p) + h_m(p) f(T_m p) \tag{3.10}$$
$$F(p) = g_m(p) + h_m(p) F(T_m p) \geq g_k(p) + h_k(p) F(T_k p),$$

whence

$$|f(p) - F(p)| \leq \text{Max} \begin{Bmatrix} |h_m(p)| |f(T_m p) - F(T_m p)| \\ |h_k(p)| |f(T_k p) - F(T_k p)| \end{Bmatrix}. \tag{3.11}$$

If we set

$$S = \sup_R |f(p) - F(p)|, \tag{3.12}$$

we obtain from (3.11) the inequality

$$|f(p) - F(p)| \leq c_2 S. \tag{3.13}$$

If we take p to be a point for which $|f(p) - F(p)| \geq S - \epsilon$, ϵ small, we obtain a contradiction, unless $S = 0$. This establishes uniqueness.

Let us observe that the uniform convergence demonstrated above establishes the further result

Theorem 2. Under the conditions

$$\text{(a)} \quad g_k(p) \text{ and } h_k(p) \text{ are continuous functions of } p \text{ in } R \tag{3.14}$$

together with the previous conditions, $f(p)$ is a continuous function of p in R.

Furthermore, if $g_k(p)$ and $h_k(p)$ are continuous functions of a set of parameters, q, $f(p)$ will be a continuous function of these parameters.

§4. Alternate Proof of Existence.

We have in the preceding section discussed the problem purely from the analytic standpoint without regard for the underlying processes. Let us now discuss the problem with regard to the basic process, and consider the process where only N stages will be allowed. If we define, similarly to (2.1), $f_N(p)$ to be maximum return for N stages, we obtain

$$f_1(p) = \max_k (g_k(p)), \tag{4.1}$$

and, generally,

$$f_{N+1}(p) = \max_k \left[(g_k(p) + h_k(p) f_N(T_k p)) \right]. \tag{4.2}$$

Let us now assume that g_k is actually a non-negative return and that $h_k(p)$ is a probability.

It is clear then that $f_2 \geq f_1$ and thus generally that $f_{N+1}(p) \geq f_N(p)$. If we set

$$U_N = \sup_R f_N(p), \tag{4.3}$$

we obtain from (4.2),

$$U_{N+1} \leq c_1 + c_2 U_N, \tag{4.4}$$

which means that $U_N \leq c_1/(1-c_2)$. Since the f_N are uniformly bounded and monotone increasing, f_N converges to $f(p)$, a solution.

§5. Approximation in Strategy Space.

The functional equation discussed in the previous section effects a transliteration of a decision problem from the space of policies, strategies, schedules, etc., to the space of functions. This is its principal role.

The essence of the previous section was that an initial guess in function space will, by the process of successive iteration, eventually yield an arbitrarily close approximation to the actual solution.

We may, however, instead of guessing an initial function, guess an initial strategy S. For example, we may divide the region R into m sub-regions, R_1, R_2, $\cdots$, R_m, possessing only boundary points in common, and choose the k^{th} choice, i.e., set

$$f_s(p) = g_k(p) + h_k(p) f_s(T_{kp}) \tag{5.1}$$

whenever $p \in R_k$. For the points on the boundary of two or more regions, we choose either index.

If $p \in R_k$, the transformed point $T_k p$ will belong to R_ℓ, where ℓ may or may not equal k. In any case, continuing in this way, we can calculate an approximation to $f(p)$ $f_s(p)$, which we can then improve by successive approximations as before.

The importance of this procedure lies in the fact that the convergence, under the assumptions of the preceding section, will always be monotone. This is of great importance in practical applications.

To show this monotonicity, let $f_2(p)$ be the second approximation. Then

$$f_2(p) = \max_{1 \le k \le m} [g_k(p) + h_k(p) f_s(T_k p)]. \tag{5.2}$$

Comparing (5.1) and (5.2), it is clear that $f_2(p) \geq f_s(p)$. From this inequality, it follows inductively that $f_{N+1}(p) \geq f_N(p)$. A further discussion, in connection with an equation of different type, will be found in [3].

§6. The Solution of (1.1).

We shall prove

Theorem 3. Consider the functional equation

$$f(x,y) = \operatorname{Max}\left\{\begin{array}{ll} A: & \sum_{k=1}^{N} p_k\left[c_k x + f(c_k' x, y)\right] \\ B: & \sum_{k=1}^{N} q_k\left[d_k y + f(x, d_k' y)\right] \end{array}\right\} \tag{6.1}$$

where

$$\begin{aligned} &(a)\quad p_k \geq 0,\ q_k \geq 0,\ \sum_{k=1}^{N} p_k < 1,\ \sum_{k=1}^{N} q_k < 1, \\ &(b)\quad 1 \geq c_k,\ d_k \geq 0,\ c_k' + c_k = d_k' + d_k = 1, \\ &(c)\quad x, y \geq 0. \end{aligned} \tag{6.2}$$

The optimal choice of operations is the following: If

$$\frac{\sum_{k=1}^{N} p_k c_k}{1 - \sum_{k=1}^{N} p_k}\, x > \frac{\sum_{k=1}^{N} q_k d_k}{1 - \sum_{k=1}^{N} q_k}\, y \tag{6.3}$$

choose A; if the reverse inequality holds, choose B. In case of equality, either choice is satisfactory.

To simplify the notation and the algebra, let us consider first the simpler form of (6.1) given by (2.6) and assume that $p_3 = q_3 = 0$. The resulting equation is

$$f(x,y) = \text{Max} \begin{cases} \text{A:} & p_1[x+f(0,y)] + p_2[r_1x+f((1-r_1)x,y)] \\ \text{B:} & q_1[y+f(x,0)] + q_2[r_2y+f(x,(1-r_2)y)] \end{cases} \tag{6.4}$$

As noted above, we already know from Section 3 that there is a unique solution to this equation. Let us turn, then, to a discussion of some of the simpler properties of $f(x,y)$. Since $p_1 + p_2 < 1$, $q_1 + q_2 < 1$, it follows that $f(0,0) = 0$. From the fact that $f(kx,ky)$ and $kf(x,y)$ satisfy the same equation for $k \geq 0$, it follows that $f(kx,ky) = kf(x,y)$, for $k \geq 0$. Setting $y = 0$ and using $f(r_1x,0) = r_1f(x,0)$, we obtain

$$\begin{aligned} f(x,0) &= \text{Max} \begin{bmatrix} \text{A:} & (p_1+p_2r_1)x + p_2(1-r_1)f(x,0) \\ \text{B:} & (q_1+q_2)f(x,0) \end{bmatrix} \\ &= (p_1+p_2r_1)x + p_2(1-r_1)f(x,0) \end{aligned} \tag{6.5}$$

whence

$$f(x,0) = \frac{(p_1+p_2r_1)x}{[1-p_2(1-r_1)]} \tag{6.6}$$

and, similarly,

$$f(0,y) = \frac{(q_1+q_2r_2)y}{[1-q_2(1-r_2)]} . \tag{6.7}$$

These results are, of course, obvious if we consider the process generating the function. On these grounds we should also

suspect that A would be employed whenever y was sufficiently small compared with x. This fact follows from the continuity of f(x,y) (compare Section 3), since the inequality

$$f(x,y) > (q_1 + q_2 r_2)y + q_1 f(x,0) + q_2 f(x,(1-r_2)y) \qquad (6.8)$$

must hold for small positive $y = y(x)$, for $x > 0$, since it is valid for $y = 0$.

It follows that there are two regions, close to the x and y axes, in which the optimal choices are, respectively, A and B, whenever (x,y) is contained in either of these regions, as shown in Fig. 2.

It is reasonable to suppose that the solution has the form shown in Fig. 1. The meaning of Fig. 1 is that A is employed whenever (x,y) is in R_A, the region between the x-axis and L, and B is employed in the complementary region. On the line L either A or B may be used.

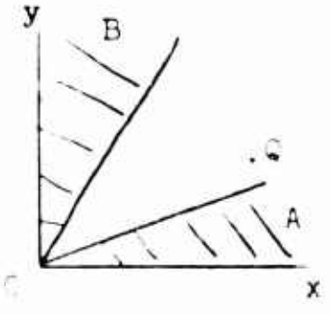

Fig. 2

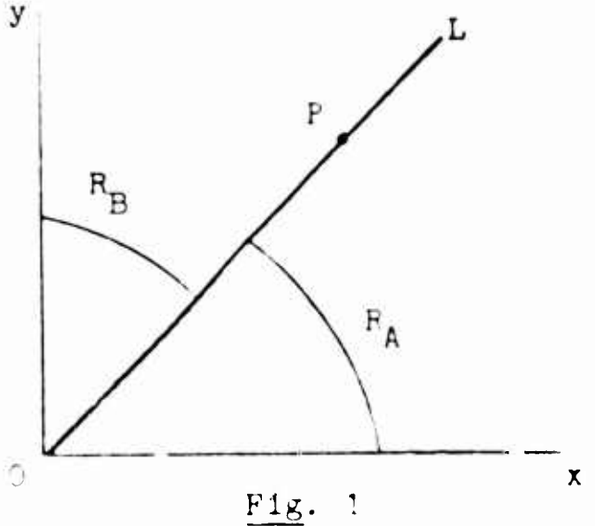

Fig. 1

That the boundary curve, if it exists, must be a straight line follows from the homogeneity of f(x,y). Assuming that the solution

has this form, we shall show that the equation of L may be calculated from the fact that it is an indifference curve. By this we mean that for points (x,y) on the curve, the value of the function $f(x,y)$ is the same whether we employ A or B.

Observe that the effect of employing A is always to drive P into R_B, whereas the use of B sends P into R_A. Consequently, if A is used at P, the next choice, in an optimal policy, must be B, and vice versa if B is used.

This alone would not be sufficient to determine L, were it not for another fact. Since the operations A and B operate on x and y alone, there will be a certain symmetry in the results obtained by using A and then B, or B and then A, which plays a decisive role in the solution.

Let us now do a small amount of computing. Using the values of $f(x,0)$ and $f(0,y)$ obtained above, we have

$$f(x,y) = \text{Max}\left[\begin{array}{ll} \text{A:} & (p_1+p_2r_1)x + \dfrac{p_1(q_1+q_2r_2)y}{[1-q_2(1-r_2)]} + p_2 f((1-r_1)x,y) \\ \text{B:} & (q_1+q_2r_2)y + \dfrac{q_1(p_1+p_2r_1)x}{[1-p_2(1-r_1)]} + q_2 f(x,(1-r_2)y) \end{array}\right] \tag{6.9}$$

To simplify the notation, let us denote the coefficients of x and y in the above equation by α_1,α_2 in A and by β_1,β_2 in B. If we employ A, we obtain, using an obvious notation,

$$f_A(x,y) = \alpha_1 x + \alpha_2 y + p_2 f((1-r_1)x,y) \tag{6.10}$$

Following this by B, we have

$$f_{AB}(x,y) = [\alpha_1 + \beta_1 p_2(1-r_1)]x + (\alpha_2 + \beta_2 p_2)y + p_2 q_2 f((1-r_1)x, (1-r_2)y). \tag{6.11}$$

Similarly, the result of B and then A is

$$f_{BA}(x,y) = (\beta_1 + q_2\alpha_1)x + [\beta_2 + q_2\beta_2(1-r_2)]y + p_2 q_2 f((1-r_1)x, (1-r_2)y). \tag{6.12}$$

If (x,y) lies upon L, we must have $f_{AB} = f_{BA}$. Equating the two expressions, we observe that the unknown function $f((1-r_1)x,(1-r_2)y)$ disappears. Consequently, we obtain for L the equation

$$[\alpha_1(1-q_2) + \beta_1(p_2(1-r_1)-1)]x = [\beta_2(q_2(1-r_2)-1)+\beta_2(1-p_2)]y \tag{6.13}$$

Using the precise values of $\alpha_1, \alpha_2, \beta_1, \beta_2$ as given by (6.9), we finally obtain, as the equation of L,

$$\frac{(p_1+p_2c_1)x}{1-p_1-p_2} = \frac{(q_1+q_2d_1)y}{1-d_1-q_2}. \tag{6.14}$$

This is a remarkably simple equation, since, as we observe, the coefficient of x deperds only on the A operation, while the coefficient of y depends only on the B operation. Furthermore, each coefficient admits of a very simple interpretation as the ratio of the expected yield of the operation to the probability of termination of the process.

Let us insert a word of warning: Although this elegant result holds for some generalizations of the functional equation, it does not hold in general, as we shall subsequently see.

Let us now prove that the solution actually has this simple form. To make the previous argument rigorous, we observe that below L, the procedure consisting of A, B, and an optimal continuation is superior to B, A, and an optimal continuation, and that the reverse is true above L. Referring to Fig. 2, let Q be a point above the known A-region, and far enough below L so that any outcome of a B-choice transforms $Q(x,y)$ into the known A-region.

To show that A is used at Q, we argue by contradiction. Suppose that B were used; then the next choice would necessarily be A. However, we have seen above that below L the procedure consisting of B, A, and an optimal continuation is inferior to A, B, and an optimal continuation. Hence, A is used at Q. It is clear that we may continue this argument until we have demonstrated that the region between L and the x-axis is an A-region. Similarly, starting from the known B-region, we may demonstrate that the region above L is a B-region.

We have carried through the proof for the simplest case of (6.1). There is no difficulty in verifying that the argument is general.

Geometrically, the pattern is as follows: When (x,y) is in R_A, A is employed until the resultant point is in R_B, at which time B is employed until the point is again in R_A, and so on.

§7. A Generalization.

There is no difficulty in extending the above analysis to the following n-dimensional equation

$$f(x_1,x_2,\cdots,x_n) = \max_i \left[\sum_{k=1}^{K} p_{ik}\left[c_{ik}x_i + f(x_1,x_2,\cdots,c'_{ik}x_i,\cdots,x_n)\right]\right] \tag{7.1}$$

where

$$\begin{aligned} &(a)\quad p_{ik} \geq 0, \quad \sum_{k=1}^{K} p_{ik} < 1, \quad i=1,2,\cdots,n, \\ &(b)\quad 1 \geq c_{ik} \geq 0, \quad c_{ik} + c'_{ik} = 1, \\ &(c)\quad x_i \geq 0. \end{aligned} \tag{7.2}$$

The decision functions are again the ratios of expected gain to probability of termination, namely,

$$D_i(x) = \frac{\sum_k p_{ik}c_{ik}}{1 - \sum_k p_{ik}}\, x_i \tag{7.3}$$

If $\max D_i(x_i)$ is attained for $i = L$, then the L^{th} choice is made unless there is equality, in which case any one of the maximizing choices is optimal.

§8. The Form of $f(x,y)$.

Having obtained a very simple characterization of the optimal policy, let us now turn our attention to the function $f(x,y)$. In general, no simple analytic representation will exist. If, however, we consider equation (5.8), which we write again as

$$f(x,y) = \text{Max} \begin{bmatrix} \alpha_1 x + a_2 y + p_2 f(c_2 x, y) \\ \beta_1 x + \beta_2 y + q_2 f(x, d_2 y) \end{bmatrix}, \qquad (8.1)$$

$(c_2 = 1 - r_1,\ d_2 = 1 - r_2)$

we shall show that if c_2 and d_2 are connected by a relation of the type $c_2^m = d_2^n$, m and n being positive integers, we shall obtain piecewise linear representations for $f(x,y)$.

It is sufficient, in order to illustrate the technique, to consider the simplest case, $c_2 = d_2$.

Let (x,y) be a point in the A-region. If A is applied, either (x,y) goes into (C,y), in which case B is used continually thereafter, or it is transformed into (c_2x,y), which may be in either an A- or a B-region. Let L_1 be the line that is transformed into L when (x,y) goes into (c_2x,y), let L_2 be the line transformed into L_1, and so on. Similarly, let M_1 be the line transformed into L when (x,y) goes into (x,d_2y), and so on. In the sector LOL_1, A is used first, followed by B, as shown in Fig. 3.

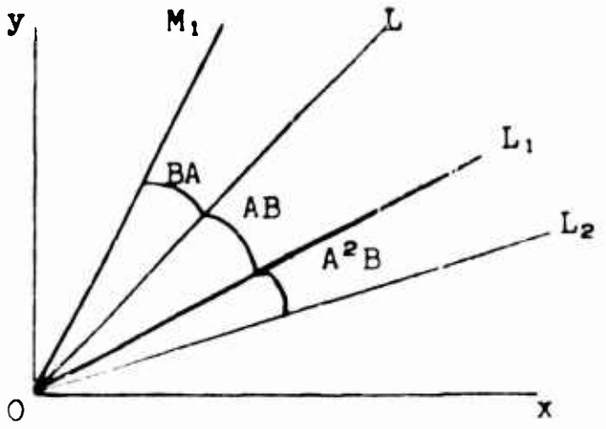

Fig. 3

Hence, for (x,y) in this sector we obtain

$$f(x,y) = a_1x + a_2y + p_2f(c_2x,y)$$
$$= a_1x + a_2y + p_2(\beta_1c_2x+\beta_2y) + p_2q_2f(c_2x,c_2y) \qquad (8.2)$$
$$= (a_1+p_2\beta_1c_2)x + (a_2+p_2\beta_2)y + p_2q_2c_2f(x,y)$$

This yields

$$f(x,y) = \frac{(a_1+p_2\beta_1c_2)x + (a_2+p_2\beta_2)y}{1 - p_2q_2c_2} \qquad (8.3)$$

for (x,y) in LOL_1. Similarly, we obtain a linear expression for f in LOM_1. Having obtained the representations in these sectors, it is clear that we obtain linear expressions in L_1OL_2, etc.

§9. The Problem for a Finite Number of Stages.

Let us now consider the problem that arises when only a finite number of stages are allowed. If we set

$$f_N(x,y) = \text{expected return using an optimal N-stage policy,} \qquad (9.1)$$

then

$$f_1(x,y) = \text{Max}\left[(p_1+p_2c_1)x,\ (q_1+q_2d_1)y\right] \qquad (9.2)$$

$$f_{N+1}(x,y) = \max \begin{cases} A: & p_1\left[x+f_N(P,y)\right] + p_2\left[c_1x+f_N(c_2x,y)\right] \\ B: & q_1\left[y+f_N(x,Q)\right] + q_2\left[d_1y+f_N(x,d_2y)\right] \end{cases}.$$

We know from the results concerning existence and uniqueness in Section 3 that, as $N \longrightarrow \infty$, $f_N(x,y) \longrightarrow f(x,y)$. However, it is not reasonable to suspect that for each N the optimal policy will be that of $f(x,y)$. Furthermore, it is clear that, in general, the policies will not be the same for $N = 1$.

It does, however, follow from our previous argumentation that if for some N the decision regions of $f_N(x,y)$ and $f(x,y)$ coincide, they must do so for all larger N.

Let us now show that decision regions for f_N converge toward that of f as $N \longrightarrow \infty$, and that there will always be an N_0 with the property that for $N \geq N_0$ the regions will coincide.

The proof is very simple. Consider the situation for $N = 2$, as in Fig. 4.

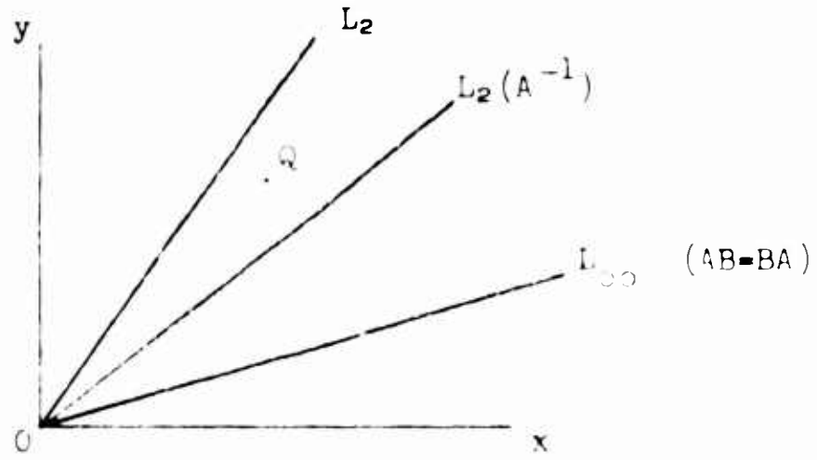

Fig. 4

Let $L_2(A^{-1})$ denote the line that is transformed into L_2 when (x,y) goes into (cx,y). Let Q be in the sector between L_2 and $L_2(A^{-1})$. If A is used at Q, then B is used next, since the transformed point is in the R_B-region for $N = 2$. If Q is above L, we

know that AB is inferior to BA, regardless of N, as a set of first two choices. Hence, B is used at x. This shows us that the B-region for the N-stage process is at least that containing the sector bounded by the y-axis and $L_2(A^{-1})$. This process continues until $L_k(A^{-1})$, for some k, lies below L_{x_0}, which must necessarily occur after some finite number of stages.

The argument is general and applies to the general equations discussed above. However, we cannot assert that the convergence is monotone, as we suspect, until we know more about the A- and B-regions for the N-stage process. It is probably true that there are two regions for each N, but this is a result that has only been demonstrated in the case of the simple equation (7.1).

To show this result, we use the fact that this equation arises from a model in which the results of an operation are known only as far as the expected outcome is concerned. Any N-stage policy has the form, therefore,

$$S_N = A^{a_1} B^{b_1} \cdots A^{a_k} B^{b_k} \tag{9.3}$$

where the a_1 and b_1 are 0 or positive integers. This notation means that the A-choice is made a_1 consecutive times, then the B-choice b_1 consecutive times, and so on. There are now two cases: S_N is either equal to A^N or B^N, or it has the form $A^k B \cdots$ or $B^\ell A \cdots$, where $k, \ell < N$.

Referring to Fig. 4, consider a point x above L. If an optimal policy has the form $A^k B \cdots, k < N$, which may be written $A^{k-1}(AB)\cdots$, it may be improved by replacing AB with BA, since A

iterated any number of times maintains Q above L. It follows then that in the region above L, either B is used first or A is used repeatedly; and, similarly, in the region below L, either A is used first or B is used repeatedly.

Since A^N is clearly the optimal policy for points sufficiently close to the x-axis, and B^N is the optimal policy for points sufficiently near the y-axis, it follows from the analytic form of the yield for any S_N—an expression which is linear in x and y—that if A^N is used at Q, it is used for all points below the line OQ, and similarly for B^N, "below" being replaced by "above."

It follows that there are always two regions, separated either by AB = BA or by a line of more complicated form, if A^N or B^N are still dominant. For large N it is clear that A^N and B^N become less and less influential, so that eventually AB = BA emerges as the sole dividing line.

§10. A General Utility Function.

We have in the previous sections considered only the case in which the utility of a total yield z was proportional to z. Let us now turn to the more interesting case in which the utility is measured by a function $\phi(z)$. The basic equation is now

$$f(x,y,a) = \text{Max} \begin{bmatrix} A: & p_1 f(0,y,a+x)+p_2 f(c_2 x,y,a+c_1 x)+p_3\phi(a) \\ B: & q_1 f(x,0,a+y)+q_2 f(x,d_2 y,a+d_1 y)+q_3\phi(a) \end{bmatrix} \tag{10.1}$$

$$f(0,0,a) = \phi(a)$$

where $c_1 + c_2 = 1$, $d_1 + d_2 = 1$, $p_1, q_1 \geq 0$,

$p_1 + p_2 + p_3 = q_1 + q_2 + q_3 = 1$.

This equation is more difficult to treat of than that occurring for $\phi(z) = z$, and we shall only be able to present its solution for certain classes of functions.

We have

$$f(0,y,a) = \text{Max} \begin{bmatrix} A: & p_1 f(0,y,a) + p_2 f(0,y,a) + p_3\phi(a) \\ B: & q_1 f(0,0,a+y) + q_2 f(0,d_2 y, a+d_1 y) + q_3\phi(a) \end{bmatrix} \tag{10.2}$$

Since $f(x,y,a) \geq f(0,0,a) = \phi(a)$ for $x,y \geq 0$, with strict inequality if x or y is positive, it follows, since $p_1+p_2+p_3=1$, $p_1 \geq 0$, that

$$f(0,y,a) = q_1\phi(a+y) + q_3\phi(a) + q_2 f(0,d_2 y, a+d_1 y) \tag{10.3}$$

and, similarly, that

$$f(x,0,a) = p_1\phi(a+x) + p_3\phi(a) + p_2 f(c_2 x, 0, a+c_1 x). \tag{10.4}$$

For given ϕ, these equations may now be solved by iteration for the functions $f(0,y,a)$ and $f(x,0,a)$.

Let us again proceed formally before turning to a justification of our operations. It is clear from the conservative nature of the processes involved that the quantity $x + y + a$ remains constant throughout the sequence of operations. Consequently, the effect of any choice is to transform a point in the region R: $x+y+a=c$, $x,y,a \geq 0$ into another point in the region, as shown in Fig. 5.

The problem that confronts us is that of determining the set of points in R in which A is used and the set in which B is used.

If we assume, as before, that these sets constitute connected regions having a boundary curve P, we may proceed to find the boundary as before, using the fact that the boundary is an indifference curve.

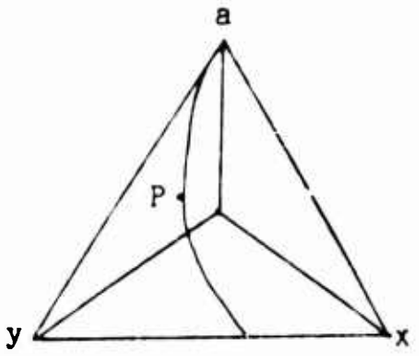

Fig. 5

However, we must assume more about the boundary curve than previously, where the fact that it was a straight line resulted in considerable simplification. Let us assume that the result of applying A to a point P on the boundary curve is to transform it into the B-region, and vice versa.

Having provided ourselves with a cushion of assumptions, let us now go through the calculations. If A is employed, we obtain

$$f(x,y,a) = p_1 f(0,y,a+x) + p_2 f(c_2x,y,a+c_1x) + p_3\phi(a) \qquad (10.5)$$

Employing B at $(0,y,a+x)$ and $(c_2x,y,a+c_1x)$, we obtain

$$\begin{aligned} f(x,y,a) = {} & p_1\left[q_1\phi(a+x+y) + q_2 f(0,d_2y,a+x+d_1y) + q_3\phi(a+x)\right] \\ & + p_2\left[q_1 f(c_2x,0,a+c_1x+y) + q_2 f(c_2x,d_2y,a+c_1x+d_1y)\right. \\ & \left. + q_3\phi(a+c_1x)\right] \\ & + p_3\phi(a). \end{aligned} \qquad (10.6)$$

A similar expression is obtained by using B and then A. Equating the two, we obtain, for the equation of the boundary curve,

$$\begin{aligned} & p_1q_3\phi(a+x) + p_2q_3\phi(a+c_1x) + p_3\phi(a) \\ = \; & q_1p_3\phi(a+y) + q_2p_3\phi(a+d_1y) + q_3\phi(a) \end{aligned} \tag{10.7}$$

which may be written

$$\begin{aligned} & p_1q_3[\phi(a+x)-\phi(a)] + p_2q_3[\phi(a+c_1x) - \phi(a)] \\ = \; & q_1p_3[\phi(a+y)-\phi(a)] + q_2p_3[\phi(a+d_1y)-\phi(a)] \end{aligned} \tag{10.8}$$

In order to establish the result rigorously, we must ascertain whether or not the boundary curve has the desired transformation property.

What we actually require is

Property T. *If*

$$\begin{aligned} F(x,y,a) = \; & p_1q_3[\phi(a+x)-\phi(a)] + p_2q_3[\phi(a+c_1x)-\phi(a)] \\ & - q_1p_3[\phi(a+y)-\phi(a)] - q_2p_3[\phi(a+d_1y)-\phi(a)] \le 0 \end{aligned} \tag{10.9}$$

then $F(c_2x,y,a+c_1x) \le 0$. *If* $F(x,y,a) \ge 0$, *then* $F(x,d_2y,a+d_1y) \ge 0$.

Unfortunately, it seems to be difficult to present any simple criterion which will insure that a general utility function $\phi(z)$ will satisfy Property T. It is not difficult to show, for example, that $\phi(z) = z^2$ does not satisfy it for all values of p_1 and q_1.

Let us now demonstrate

Theorem 4. If

(a) $\phi(z)$ is strictly increasing and continuous,

$\phi(z) \geq 0$, (10.10)

(b) Property T is satisfied,

then the solution to (10.1) is given by

$$f(x,y,a) = p_1 f(0,y,a+x) + p_2 f(c_2 x, y, a+c_1 x) + p_3 \phi(a) \quad (10.11)$$

for $F(x,y,a) \geq 0$, and by

$$f(x,y,a) = q_1 f(x,0,a+y) + q_2 f(x,d_2 y, a+d_1 y) + p_3 \phi(a) \quad (10.12)$$

for $F(x,y,a) \leq 0$.

The optimal policy is to apply A when $F(x,y,a) > 0$ and B if $F(x,y,a) < 0$. When there is equality, it is a matter of indifference as to which choice is made.

Proof: The proof is carried through in two stages. First we show that there is a region in the plane $x+y+a=c$ where A is always used, namely, a region close to $y=0$. Then we consider what happens at a point Q in the region defined by $F(x,y,a) \geq 0$ and $x+y+a=c$.

Let us assume for the moment that we have already established the existence of a region where A is always used. If B is used at Q, it follows from Property T that the transformed point is again in the same region. It cannot be true that B is used repeatedly if $x > 0$, since eventually the y coordinate will be so small that

the point will be in the A-region. Hence, if at x an optimal policy employs B for the first k choices, the sequence of moves has the form

$$S = BB\cdots(k \text{ times})\cdots BA. \tag{10.13}$$

On the basis of Property T, we are still in the region $F(x,y,a) \geq 0$, $x+y+a=c$ after employing B (k-1) times. The next two moves, B and then A, cannot be optimal, however, since the region is defined by the property that AB plus optimal continuation is superior to BA plus optimal continuation. This shows that at x, move B cannot be used first in an optimal policy.

It remains then to establish the existence of the A-region mentioned above. Since $f(x,y,a) > \phi(a)$ for $x,y \geq 0$ and one at least positive, it follows that

$$\begin{aligned} &p_1 f(0,y,a+x) + p_2 f(c_2 x, y, a+c_1 x) + p_3 \phi(a) \\ &> q_1 f(x,0,a+y) + q_2 f(x, d_2 y, a+d_1 y) + q_3 \phi(a) \end{aligned} \tag{10.14}$$

which holds at $y=0$, must by virtue of the continuity of the functions involved, for any $x > 0$, hold for some interval $0 \leq y \leq y(x,a)$.

§4. The Exponential Utility Function.

One way of obtaining utility functions that have the desired property is to make the boundary equation independent of a. If we wish this to be true for all values of the parameters p_1 and q_1, we must have

$$\phi(a+x) - \phi(a) = G(x)H(a) \tag{11.1}$$

which yields, using standard arguments, under the assumption of continuity,*

$$\begin{aligned} &(a)\quad \phi(z) = mz+n \quad \text{or} \\ &(b)\quad \phi(z) = ce^{bz}\ . \end{aligned} \tag{11.2}$$

We have already considered the first utility function: let us now consider the second.

The important property of these utility functions is that a policy which maximizes the expected value of $\phi(z)$ proceeds at each stage without regard for the amount already obtained, being dependent only on the remaining amount to be obtained.

If we set, for $b > 0$,

$$g(x,y) = \text{Max Exp } (e^{bz}) \tag{11.3}$$

("Exp" denoting here "expected value," not "exponential"), we obtain for g the functional equation

$$g(x,y) = \text{Max} \begin{bmatrix} \text{A:} & p_1 e^{bx} g(0,y) + p_2 e^{bc_1 x} g(c_2 x, y) + p_3 \\ \text{B:} & q_1 e^{by} g(x,0) + q_2 e^{bdy} g(x, d_2 y) + q_3 \end{bmatrix}. \tag{11.4}$$

As a special case of Theorem 4, we obtain

Theorem 5. The solution of (11.4) is as follows: For

$$\frac{p_1(e^{bx}-1) + p_2(e^{bc_1 x}-1)}{p_3} > \frac{q_1(e^{by}-1) + q_2(e^{bd_1 y}-1)}{q_3} \tag{11.5}$$

* This requirement of continuity can be considerably weakened.

use A; if the reverse inequality holds, employ B; if equal, either is applicable.

Observe that, as should be true, the limiting solution as $b \longrightarrow 0$ is exactly that obtained from $\phi(x) \equiv z$.

§12. Asymptotic Behavior of $g(x,y)$.

We now turn to the problem of determining the asymptotic behavior of $g(x,y)$ as x and $y \longrightarrow \infty$. We begin by deriving the asymptotic behavior of $g(x,0)$ and $g(0,y)$. From the equation we obtain, for large x,

$$g(x,0) = p_1 e^{bx} + p_3 + p_2 e^{bc_1 x} g(c_2 x, 0). \tag{12.1}$$

This equation may be solved by iteration:

$$g(x,0) = (p_1 e^{bx} + p_3) + p_2 e^{bc_1 x}(p_3 + p_1 e^{bc_2 x}) \tag{12.2}$$

To obtain the asymptotic behavior, however, we must proceed differently. Set

$$g(x,0) = \frac{p_1 e^{bx}}{1-p_2} + h(x) e^{bx} \tag{12.3}$$

where h satisfies the equation

$$h(x) = p_3 e^{-bx} + p_2 h(c_2 x) \tag{12.4}$$

as we see by direct substitution. Although iteration yields

$$h(x) = p_3 e^{-bx} + p_2 p_3 e^{-bc_2 x} + \cdots \qquad (12.5)$$

the asymptotic behavior of $h(x)$ is still not apparent. We shall show that $h(x) = x^{-2}\psi(x)[1+o(1)]$ as $x \longrightarrow \infty$, where $\psi(x) = \psi(c_2 x)$, $a = (\log 1/p_2)/(\log 1/c_2)$. To accomplish this, set $h(x) = k(x)x^{-a}$. Then k satisfies the simpler equation

$$k(x) - k(c_2 x) = p_3 x^a e^{-bx} = \phi(x). \qquad (12.6)$$

The essential fact about ϕ that we shall use is that $\sum_{n=1}^{\infty} \phi(x/c_2^n)$ converges for each x. From (12.6) we have

$$k(x/c_2^n) - k(x/c_2^{n-1}) = \phi(x/c_2^n) \qquad (12.7)$$

which yields

$$\lim_{n \longrightarrow \infty} k(x/c_2^n) = k(x) + \sum_{n=1}^{\infty} \phi(x/c_2^n) = \psi(x). \qquad (12.8)$$

From the form of the limit function or from the equation for $k(x)$, we see that $\psi(x) = \psi(c_2 x)$ for all x. If then we write $y = x/c_2^n$ for $1 \leq x \leq 1/c_2$, we have

$$k(y) - k(x/c_2^n) = [1+o(1)]\psi(x) = [1+o(1)]\psi(x/c_2^n) \qquad (12.9)$$

as $y \longrightarrow \infty$.

Collecting the previous results, we see that the asymptotic behavior of $g(x,0)$ is given by

$$g(x,0) = \frac{p_1 e^{bx}}{1-p_2} + \frac{e^{bx}\phi(x)}{x^{a_1}}\left[1 + o(1)\right] \tag{12.10}$$

where

$$\begin{aligned} &(a)\quad \phi(x) = \phi(c_2 x) \\ &(b)\quad a_1 = \frac{\log 1/p_2}{\log 1/c_2} \end{aligned} \tag{12.11}$$

The corresponding result for $g(0,y)$ is

$$g(0,y) = \frac{q_1}{1-q_2} e^{by} + \frac{e^{by}\delta(y)}{y^{b_1}}\left(1 + o(1)\right) \tag{12.12}$$

where

$$\begin{aligned} &(a)\quad \delta(y) = \delta(d_2 y) \\ &(b)\quad b_1 = \log 1/q_2 / \log 1/d_2 \end{aligned} \tag{12.13}$$

Turning to the equation for $g(x,y)$, we have for x and y large

$$g(x,y) = \text{Max}\left[\begin{array}{l} \frac{p_1 q_1}{1-q_2} e^{b(x+y)} + p_2 e^{bc_1 x} g(c_2 x, y) + O(e^{by}/y^{b_1}) \\ \\ \frac{p_1 q_1}{1-p_2} e^{b(z+y)} + q_2 e^{bd_1 y} g(x, d_2 y) + O(e^{bx}/x^{a_1}) \end{array}\right] \tag{12.14}$$

Setting $h(x,y)^{b(x+y)} = g(x,y)$, we obtain

$$h(x,y) = \text{Max}\left[\begin{array}{l} \frac{p_1 q_1}{1-q_2} + p_2 h(c_2 x, y) + O(e^{-bx} y^{-b_1}) \\ \\ \frac{p_1 q_1}{1-p_2} + q_2 h(x, d_2 y) + O(e^{-by} x^{-a_1}) \end{array}\right] \tag{12.15}$$

To simplify still further, we set $h(x,y) = \alpha + k(x,y)$, obtaining

$$\alpha+k(x,y) = \text{Max} \begin{bmatrix} \frac{p_1 q_1}{1-q_2} + \alpha p_2 + p_2 k(c_2 x, y) + O(e^{-bx_y - b_1}) \\ \\ \frac{p_1 q_1}{1-p_2} + \alpha q_2 + q_2 k(x, d_2 y) + O(e^{-by_x - a_1}) \end{bmatrix} \tag{12.16}$$

If α is chosen to be the common solution of

$$\alpha = \frac{p_1 q_1}{1-q_2} + p_2 \alpha = \frac{p_1 q_1}{1-p_2} + q_2 \alpha \tag{12.17}$$

namely, $p_1 q_1/(1-p_1)(1-q_1)$, (12.16) simplifies to

$$k(x,y) = \text{Max} \begin{bmatrix} p_2 k(c_2 x, y) + O(e^{-bx_y - b_1}) \\ q_2 k(x, d_2 y) + O(e^{-by_x - a_1}) \end{bmatrix} \tag{12.18}$$

To estimate $k(x,y)$ we use the fact that the solution may be obtained by means of successiveapproximations:

$$k_{n+1}(x,y) = \text{Max} \begin{bmatrix} p_2 k_n(c_2 x, y) + O(e^{-bx_y - b_1}) \\ \\ q_2 k_n(x, d_2 y) + O(e^{-by_x - a_1}) \end{bmatrix}, \quad k_0(x,y) = 1/x^r + y^r , \tag{12.19}$$

considering, for our purposes, only values of x and y greater than 1. The exponent r will be chosen in a moment.

If we have an inequality of the type $k_n(x,y) \leq u_n/(x^r+y^r)$, u_n being a constant, which inequality is certainly valid for n=0, we obtain

$$k_{n+1}(x,y) \leq \text{Max} \left[\begin{array}{l} \dfrac{p_2 u_n}{c_2^r(x^r+y^r)} + O(e^{-bx}y^{-b_1}) \\ \\ \dfrac{}{d_2^r(x^r+y^r)} + O(e^{-by}x^{-a_1}) \end{array} \right] \tag{12.20}$$

Choose r so that $p_2 c_2^{-r} \leq 1/2$, $q_2 d_2^{-r} \leq 1/2$. Since $a_1, b_1 > r$, we see, since $x^a e^{-bx} \leq d_r$ for all x, that $e^{-bx}y^{-a_1} \leq d_r x^{-r}y^{-r} \leq d_r/(x^r+y^r)$, for $x,y \geq 1$. Hence, we have

$$k_{n+1}(x,y) \leq \text{Max} \left[\begin{array}{l} \dfrac{\frac{1}{2} u_n}{x_r+y^r} + \dfrac{a_2}{x^r+y^r} \\ \\ \dfrac{\frac{1}{2} u_n}{x^r+y^r} + \dfrac{a_2}{x^r+y^r} \end{array} \right] \tag{12.21}$$

for some constant a_2. If we take $u_{n+1} = \frac{1}{2}(u_n+a_2)$, the inequality is preserved for u_{n+1}. Since u_n as defined by the recurrence relation is uniformly bounded, we obtain, in the limit, $k(x,y) \leq a_3/(x^r+y^r)$.

Knowing the form of the function, we readily obtain the optimal policy, deriving in this case the slightly paradixocal result that, asymptotically, as x and y $\longrightarrow \infty$, it makes no difference which move is made first.

Collecting the above results, we obtain

$$g(x,y) = \frac{e^{b(x+y)}p_1 q_1}{(1-p_2)(1-q_2)} + O(e^{b(x+y)}/x^r+y^r \tag{12.22}$$

§13. A Continuous Version.

As we have seen in the previous sections, the formulation of the gold-mining problem in its discrete form leads to a number of unsolved problems. We turn, therefore, to a continuous version of the problem in the hope of overcoming our difficulties by use of the more powerful tools of continuity. We can now resolve the corresponding questions in complete detail and thereby obtain a clear insight into the structure of the optimal policies. The solutions determined in this way can now be used as approximations in the original discrete process.

One very interesting and crucial fact emerges. Whereas the original discrete problem had certain linear aspects, at least in the case where we were considering expected return, the continuous version is sufficiently nonlinear to permit a variational approach in the classical manner. In carrying through this variational attack our knowledge of the form of the solution in the discrete formulation is of great service in telling us in advance what to expect. It is a combination of the two techniques, old and new, which permit a successful attack on the problem.

Let us now begin by discussing some methods we may follow to obtain a continuous analogue to (1.1). The basic assumption is that each operation is to have a high probability of obtaining a small amount and leaving the machine undamaged, and a small probability of obtaining nothing and damaging the machine.

Let, for $\delta > 0$ and small,

$1 - q_1\delta$ = probability of obtaining $r_1x\delta$ and leaving the machine undamaged, if A is used

$1 - q_2\delta$ = probability of obtaining $r_2y\delta$ and leaving the machine undamaged if B is used, (13.1)

where $q_1, q_2 > 0$.

Setting, as before, $f(x,y)$ equal to the total expected gain obtained before the machine is damaged, we obtain the functional equation

$$f(x,y) = \text{Max} \begin{bmatrix} A: & (1-q_1\delta)(r_1x\delta + f(x-r_1x\delta, y) \\ B: & (1-q_2\delta)(r_2y\delta + f(x, y - r_2y\delta) \end{bmatrix} \tag{13.2}$$

If we proceed formally, letting $\delta \longrightarrow 0$, we obtain

$$f(x,y) = \text{Max} \begin{bmatrix} A: & f(x,y) + r_1x\delta - q_1\delta f(x,y) - r_1x\delta\, f(x,y) + O(\delta^2) \\ B: & f(x,y) + r_2y\delta - q_2\delta f(x,y) - r_2y\delta\, f(x,y) + O(\delta^2) \end{bmatrix} \tag{13.3}$$

or

$$0 = \text{Max} \begin{bmatrix} A: & r_1x - q_1 f(x,y) - r_1x\, f_x(x,y) \\ B: & r_2y - q_2 f(x,y) - r_2y\, f_y(x,y) \end{bmatrix}. \tag{13.4}$$

This does not seem to be a fruitful approach because of the difficulty of establishing any existence or uniquenss theorems, or, in general, of treating the equation in (13.4) analytically.

In place of using a differential approach, we may use an integral approach and then let $\delta \longrightarrow 0$. Let us use (13.2) and iterate, obtaining the corresponding equation for n steps at a time. The result has the form

$$f(x,y) = \operatorname*{Max}_{S_n} \left[R_n(x,y) + \sum_k p_{nk}(x,y) f(x_{nk}, y_{nk}) \right] \tag{13.5}$$

where

$R_n(x,y)$ = expected return from n stages using the policy S_n, (13.6)

$P_{nk}(x,y)$ = probability of surviving and being at (x_{nk}, y_{nk}) using S_n,

S_n = policy pursued; i.e., the choice of A or B at each stage.

If $n\delta$ is chosen to remain finite as $\delta \longrightarrow 0$, $n \longrightarrow \infty$, and set equal to t, the analogue of (13.5) is a functional equation of the type

$$f(x,y) = \operatorname*{Max}_{S(t)} \left[R(x,y,t) + \int_{r=0}^{1} \int_{s=0}^{1} f(xr, ys)\, dG_S(r,s,x,y,t) \right] \tag{13.7}$$

Functional equations of this class will occur in most continuous versions of dynamic programming problems. We shall not enter into any discussion of this formulation here because of the many

conceptual and mathematical difficulties associated with the concept of a continuous strategy, particularly when the outcome is stochastic. Instead we shall use a third approach which bears the same connection to (13.7) as the use of the heat equation in diffusion theory bears to the Chapman-Kolmogoroff equation. At the moment it is sufficient for our purposes.

Let us begin by noting that according to the results of §6 the solution of (13.2) is determined by the boundary curve

$$\frac{(1-q_1\delta)r_1x\delta}{q_1\delta} = \frac{(1-q_2\delta)r_2y\delta}{q_2\delta}, \tag{13.8}$$

which as $\delta \longrightarrow 0$ approaches the line

$$r: \quad r_1x/q_1 = r_2y/q_2. \tag{13.9}$$

The following strategy is the limit of the strategies as $\delta \longrightarrow 0$:

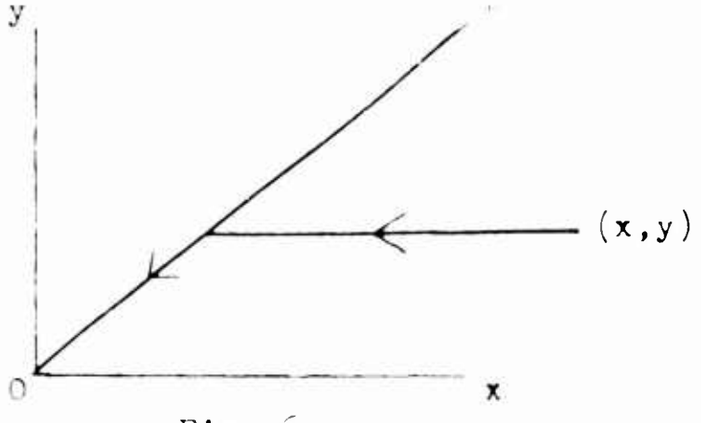

Fig. 6

If (x,y) is below r, use A, continuing across horizontally until r is hit, and then continuing down r.

A strategy of this type is not included in the original formulation of the problem which allowed only horizontal or vertical

motion, i.e., use of A or B. It is clear, however, that this policy can be arbitrarily closely approximated by use of A and B moves. This suggests that the continuous version of the original problem may not possess a policy yielding a maximum return, but only a sequence of policies yielding a supremum.

However, the introduction of mixing at a point introduces a number of difficulties of both physical and mathematical kind. Mathematically, we find ourselves confronted by the same difficulties that made us wish to bypass (13.7); physically, we are reluctant to accept this type of policy as one applicable to the original problem which insisted upon a choice of A or B.

To avoid this concept of mixing at a point, we use a frequently useful device. The essence of it is that for mathematical purposes mixing over small intervals is equivalent to mixing at a point itself, under certain natural continuity assumptions; cf. [6] for a further discussion.

We shall assume then that we are considering a process which requires a choice of A or B at time points $0,\Delta,2\Delta,\cdots$, etc., and that over a typical interval $[k\Delta,(k+1)\Delta]$ we use A for the part $[k\Delta,k\Delta+\phi_1\Delta]$ and B for $[k\Delta+\phi_1\Delta,(k+1)\Delta]$, where ϕ_1 depends upon k.

Assuming that Δ is small and that first-order terms are sufficient to describe the process, we shall derive a set of differential equations which determine the process.

Having set up the equation, we shall, to illustrate the power of the method, solve in turn problems corresponding to the two-choice, three-choice, two-choice finite time, and general nonlinear utility function.

§14. Derivation of the Differential Equation.

We assume as above that the total time interval is divided into small intervals of length Δ. In a typical interval $[k\Delta,(k+1)\Delta]$ the first part of the interval $[k\Delta,k\Delta+\phi_1\Delta]$ is devoted to the use of A.

If x is the amount of gold in mine A at the time $k\Delta$, there is a probability $1 - q_1\phi_1\Delta$ that an amount $r_1x\phi\Delta$ is mined and the operation may be continued; and a probability $q_1\phi_1\Delta$ that nothing is obtained and the operation stops. The second part of the interval $[k\Delta+\phi_1\Delta,(k+1)\Delta]$ is devoted to the use of B. If mine B contains an amount y, then there is a probability $1 - q_2\phi_2\Delta$ that the amount $r_2y\phi_2\Delta$ is obtained and the operation may be continued; and a probability $q_2\phi_2\Delta$ that the operation ceases, where $\phi_2 = 1 - \phi_1$.

As far as first-order terms in Δ are concerned, it makes no difference in what order the operations are performed. It is this feature which allows this type of mixing to perform the function of mixing at a point.

A strategy consists of a choice of ϕ_1 and ϕ_2 for each interval. For any given strategy, let

$x(t)$ = amount of gold remaining in A provided the operation has continued to t,

$y(t)$ = amount of gold remaining in B provided the operation has continued to t,

$p(t)$ = probability that the machine survives until t, i.e., that the operation continues until t.

$f(t)$ = expected amount of gold mined up to time t, (14.1)

where $t = n\Delta$, $n=0,1,2,\cdots$.

Ignoring the second-order terms in Δ, we have

$$\begin{aligned} x(t+\Delta) &= x(t) - r_1\phi_1(t)x(t)\Delta \\ y(t+\Delta) &= y(t) - r_2\phi_2(t)y(t)\Delta \\ p(t+\Delta) &= p(t)(1 - q_1\phi_1(t)\Delta - q_2\phi_2(t)\Delta) \\ f(t+\Delta) &= f(t) + p(t)\,[\phi_1(t)r_1x(t) + \phi_2(t)r_2y(t)]\Delta . \end{aligned} \tag{14.2}$$

Letting $\Delta \longrightarrow 0$, we obtain the system of differential equations

$$\begin{aligned} dx/dt &= -\phi_1(t)r_1x(t), & x(0) &= x_0, \\ dy/dt &= -\phi_2(t)r_2y(t), & y(0) &= y_0, \\ dp/dt &= -p(t)\,[\phi_1(t)q_1 + \phi_2(t)q_2], & p(0) &= 1, \\ df/dt &= p(t)\,[\phi_1(t)r_1x(t) + \phi_2(t)r_2y(t)], & f(0) &= 0. \end{aligned} \tag{14.3}$$

The problem is now to determine ϕ_1, where

$$0 \le \phi_1(t) \le 1, \quad \phi_2(t) = 1 - \phi_1(t), \tag{14.4}$$

so as to maximize $f(T)$. A case of particular importance is $T = \infty$. We shall derive similar equations for the three-choice problem.

§15. The Variational Procedure.

Let ϕ_1 and ϕ_2 be functions furnishing the maximum, and let

$$\phi_1 = \phi_1 + \varepsilon\beta_1(t) \tag{15.1}$$

where ε is a small positive quantity, β_1, β_2 are two functions satisfying the conditions

$$0 \le \phi_1 + \varepsilon\beta_1 \le 1, \tag{15.2}$$

(which means $|\eta_1| \leq 1/\varepsilon$), and $\eta_1+\eta_2=0$, so that the ϕ_1 are also admissible ϕ's.

It follows that $\eta_1(t) \leq 0$ if $\phi_1(t) = 1$, $\eta_1(t) \geq 0$ if $\phi_1(t) = 0$, and η_1 can be of either sign of $0 < \phi_1(t) < 1$, the region where free variation is permitted. Performing the variation, we find readily that

$$\bar{x}(t) = x(t)(1-\varepsilon r_1 B_1(t)) + o(\varepsilon) \tag{15.3}$$

$$\bar{y}(t) = y(t)(1-\varepsilon r_2 B_2(t)) + o(\varepsilon),$$

$$\bar{p}(t) = p(t)(1-\varepsilon\eta_1 B_1(t)-\varepsilon\eta_2 B_2(t)) + o(\varepsilon)$$

$$\bar{f}(T) - f(T) = \int_0^T \Big\{ -f'(t)(\eta_1 B_1(t)+\eta_2 B_2(t)) + r_1 B_1(t)p(t)x'(t)$$
$$+r_2 B_2(t)p(t)y'(t) + r_1\eta_1(t)p(t)x(t)$$
$$+ r_2\eta_2(t)p(t)y(t)\Big\} dt + o(\varepsilon)$$

where we have set

$$B_1(t) = \int_0^t \eta_1(s)ds, \tag{15.4}$$

and the bars refer to the perturbed variables.

Integrating by parts to eliminate the $B_1(t)$, we find

$$\bar{f}(T) - f(T) = \varepsilon \int_0^T \left[K_1(t)\eta_1(t) + K_2(t)\eta_2(t)\right] dt + o(\varepsilon), \tag{15.5}$$

where

$$K_1(t) = -\eta_1 \int_t^T f'(s)ds + r_1 p(T)x(T) - r_1 \int_t^T p'(s)x(s)ds$$
$$\tag{15.6}$$
$$K_2(t) = -\eta_2 \int_t^T f'(s)ds + r_2 p(T)y(T) - r_2 \int_t^T p'(s)y(s)ds.$$

Since $\bar{f}(T) - f(T) \leq 0$, we see that whenever $K_1(t) > K_j(t)$ we must have $\phi_1(t) = 1$, $\phi_j(t) = 0$. These relations yield implicit equations for ϕ_1 and ϕ_j. In the next section we shall discuss the behavior of the K-functions in more detail.

§16. The Behavior of K_1.

The fundamental relation is

$$\frac{d}{dt}(K_1-K_2) = (q_1-q_2)f'(t) - p'(t)(r_2y-r_1x)$$
$$= p\,[\,q_1r_2y - q_2r_1x\,]. \tag{16.1}$$

Thus a "mixed policy," one for which more than one of the ϕ_1 is positive for a given t, which implies $K_1(t) = K_2(t)$, can be optimal only on the line $q_1r_2y = q_2r_1x$. This line is precisely the boundary line that one obtains by passage to the limit from the solution in the discrete case as $\Delta \longrightarrow 0$, as in (12.9).

If a mixed policy is pursued along the line, ϕ_1 and ϕ_2 must be chosen to stay on this line, which means that the slope $s = y/x$ must be kept constant. Since

$$\frac{d}{dt}(y/x) = \frac{y'}{x} - \frac{x'}{x}s(t) = [\,r_1\phi_1 - r_2\phi_2\,]\,s, \tag{16.2}$$

we see that we must have

$$\phi_1 = \frac{r_2}{r_1+r_2}, \quad \phi_2 = \frac{r_1}{r_1+r_2} \tag{16.3}$$

§17. The Solution for $T = \infty$.

With these preliminaries out of the way, let us determine the optimal policy for the infinite process, $T = \infty$. The infinite problem is, as usual, simpler than the finite case because of the homogeneity introduced by infinite time; after any initial actions, we are confronted by a problem of the same type, with different initial values. Let us note that a consequence of this is the homogeneity of the equations with respect to x and y is that the decision at any point is a function only of the slope $s = y/x$.

Let us begin by observing that above the line $q_1 r_2 y = q_2 r_1 x$ in the (x,y) plane if policy A is ever used, it is used thereafter. This follows immediately from (1) of §16 which shows that $K_1 - K_2$ is increasing when $q_1 r_2 y - q_2 r_1 x > 0$. Since use of A decreases x and leaves y unchanged, once $K_1 > K_2$ the use of A maintains the inequality.

Near the y-axis, however, the use of A continually is not as rewarding as continual use of B. For if $\phi_1 = 1$, $\phi_2 = 0$, for $t \geq 0$, we have

$$
\begin{aligned}
x(t) &= x_0 e^{-r_1 t} \\
y(t) &= y_0 \qquad\qquad (17.1) \\
p(t) &= e^{-q_1 t} \\
f(t) &= \int_0^t r_1 x_0 e^{-r_1 s}\, e^{-q_1 s}\, ds \\
f_A(\infty) &= r_1 x_0/(q_1 + r_1)
\end{aligned}
$$

However, $\phi_1 = 0$, $\phi_2 = 1$ for all t yields similarly $f_B(\infty)$ $= r_2 y_0/(q_2+r_2)$. For y_0/x_0 sufficiently large $f_B(\infty) > f_A(\infty)$. Thus, there is a region near the y-axis where B is used.

This region where B is used extends down to the line $q_1 r_2 y = q_2 r_1 x$. To prove this we observe that a mixed policy cannot be pursued above the line and if A is ever used above the line it is always used thereafter. Using A indefinitely, however, would eventually take (x,y) into the region near the y-axis where B is known to be optimal, a contradiction. Hence B is always used above the line. Similarly, below the line A is always used.

When the line $q_1 r_2 y = q_2 r_1 x$ is reached, the point (x,y) must remain on the line thereafter. For if not, then an A policy must be used in a B region or vice versa, which is impossible. Hence, on the line itself the mixed policy of (14.3) must be employed.

We have thus demonstrated

Theorem 6. With reference to the equations (14.3) and the constraints (14.4), the maximum value of $f(\infty)$ is attained by use of the policy

$$\phi_1 = 1 \text{ for } q_1 r_2 y < q_2 r_1 x,$$
$$\phi_2 = 1 \text{ for } q_1 r_2 y > q_2 r_1 x, \qquad (17.2)$$
$$\phi_1 = \frac{r_2}{r_2+r_2}, \quad \phi_2 = \frac{r_1}{r_1+r_2} \text{ for } q_1 r_2 y = q_2 r_1 x.$$

Note that ϕ_1 and ϕ_2 are determined almost everywhere by the above arguments, and hence are essentially unique.

§18. Solution for Finite Total Time.

In finding the solution for finite T, we shall begin by determining what policy is used last. Since an optimal policy has the property that its continuation after any initial part is also optimal, we shall consider the case where T is small. We have, for T close to 0,

$$f(T) = \int_0^T p(s)\left[\phi_1(s)r_1x(s) + \phi_2(s)r_2y(s)\right]ds$$
$$= r_1x_0\int_0^T \phi_1(s)ds + r_2y_0\int_0^T \phi_2(s)ds + o(T) \tag{18.1}$$

It follows then that for small T the maximum is obtained by taking $\phi_1(s) = 1$, $\phi_2(s) = 0$ for $r_1x > r_2y$ and $\phi_1(s) = 0$, $\phi_2(s) = 1$ for $r_2y > r_1x$. As is to be expected, for small durations expected gain, without worry about termination, is the determining factor.

If $q_1 = q_2$ the lines $r_2y = r_1x$ and $q_1r_2y = q_2r_1x$ coincide, and the optimal policy is easily found to be the same as that for $T = \infty$.

Let us consider the general case where $q_1 \neq q_2$. Assume, without loss of generality, that the line $r_2y = r_1x$ lies above the line $q_1r_2y = q_2r_1x$. The positive quadrant is then divided into three regions, which we label I, II, III.

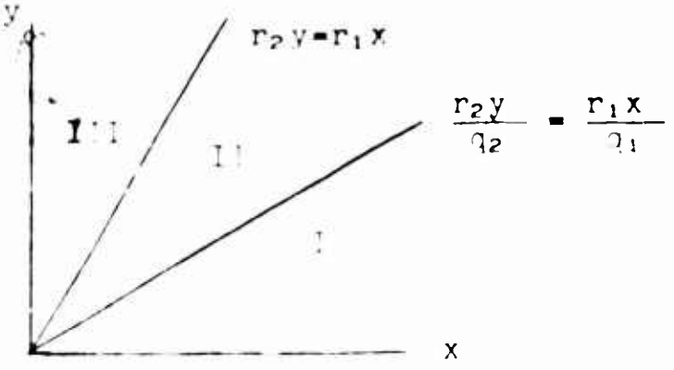

Fig. 7

As before, it follows that in region I a B-policy once used must be continued thereafter, while in regions II and III the same holds for an A-policy. Also, in regions I and II an A-policy is used if the time resulting is sufficiently small, and in III a B-policy under the same conditions. From this we conclude that an A-policy is always used in I, and a B-policy always while in III.

Let us now establish that an optimal policy never switches from A to B. Let us suppose otherwise and let t_o be the time at which the change occurs. Since at t_o, A is terminated, the point $(x(t_o),y(t_o))$ must be in region I, or on the boundary between I and II. Using B will keep the point $(x(t),y(t))$ in I for all $t > t_o$ since we know that B once used in I must be continued. However, this contradicts the fact that A is used in I whenever the time remaining is sufficiently small. Similarly, the combination of using the mixed policy and then B cannot occur, since the change-over must occur on the boundary between I and II, and B used thereafter in region I, a contradiction.

This reduces the number of types of solutions to six: A always, B always; the mixed policy followed by A; A then the mixed policy and finally A; B then the mixed policy and then A; B followed by A.

Let t_o be the value of t at which the last change of policy is made in an optimal strategy, if such a change occurs. For $t_o < t \leq T$, we must have $\phi_1(t) = 1$, $\phi_2(t) = 0$. We now compute the value of $K_1(t_o) - K_2(t_o)$. We have for $t_o < t \leq T$,

$$x(t) = x(t_o)e^{-r_1(t-t_o)}, \quad y(t) = y(t_o)$$

$$p(t) = p(t_o)e^{-q_1(t-t_o)}, \tag{18.2}$$

$$f'(t) = p(t_o)e^{-(q_1+r_1)(t-t_o)}r_1x(t_o),$$

and, after some simplification,

$$K_1(t_o) - K_2(t_o) = p(t_o)r_1x(t_o)\left[\left(1 - \frac{q_2}{q_1+r_1}\right)e^{-(q_1+r_1)(T-t_o)} + \frac{q_2}{q_1+r_1} - \frac{r_2y(t_o)}{r_1x(t_o)}\right] \tag{18.3}$$

For any fixed point $(x(t_o),y(t_o))$ in II, the right side is positive for $T-t_o$ small, and negative for $T-t_o$ large. It is equal to zero for precisely one value of $T-t_o$. This zero determines when the changeover occurs. When it occurs, A is used for the remaining time, with any of the six beginnings above, depending upon the location of the initial point.

§19. The Three-choice Problem.

The continuous version of the three-choice problem mentioned above is the following: Given

$$\begin{aligned}
&\frac{dx}{dt} = -\left[\phi_1(t)r_1 + \phi_3(t)r_3\right]x(t), && x(0) = x_o\\
&\frac{dy}{dt} = -\left[\phi_2(t)r_2 + \phi_3(t)r_4\right]y(t), && y(0) = y_o \qquad (19.1)\\
&\frac{dp}{dt} = -p(t)\left[\phi_1(t)q_1 + \phi_2(t)q_2 + \phi_3(t)q_3\right], && p(0) = 1,\\
&\frac{df}{dt} = p(t)\left[(\phi_1(t)r_1 + \phi_3(t)r_3)x(t) + (\phi_2(t)r_2 + \phi_3(t)r_4)y(t)\right], && f(0) = 0
\end{aligned}$$

where for all t

$$0 \le \phi_i \le 1, \quad \phi_1 + \phi_2 + \phi_3 = 1, \tag{19.2}$$

it is required to determine the $\phi_1(t)$ so as to maximize $f(T)$.

We shall consider only the case where $T = \infty$.

As before, let us set $\bar{\phi}_1 = \phi_1 + \beta_1$, and $B_1(t) = \int_0^t \beta_1(s)ds$.

We obtain

$$\bar{x}(t) = x(t)(1 - \varepsilon r_1 B_1(t) - \varepsilon r_3 B_3(t)) + o(\varepsilon)$$

$$\bar{y}(t) = y(t)(1 - \varepsilon r_2 B_2(t) - \varepsilon r_4 B_3(t)) + o(\varepsilon) \tag{3}$$

$$\bar{p}(t) - p(t)(1 - \varepsilon \sum_{i=1}^{3} q_i B_i(t)) + o(\varepsilon)$$

$$\frac{d\bar{f}}{dt} = \bar{p}\left[(\bar{\phi}_1 r_1 + \bar{\phi}_3 r_3)\bar{x} + (\bar{\phi}_2 r_2 + \bar{\phi}_3 r_4)\bar{y} \right]$$

Consequently, following the same technique as before, we obtain

$$\bar{f}(T) - f(T) = \varepsilon \int_0^T \left[K_1\beta_1 + K_2\beta_2 + K_3\beta_3 \right] dt + o(\varepsilon) \tag{4}$$

where

$$K_1(t) = -q_1 \int_t^T f'(s)ds + r_1 p(T)x(T) - r_1 \int_t^T p'(s)x(s)ds$$

$$K_2(t) = -q_2 \int_t^T f'(s)ds + r_2 p(T)y(T) - r_2 \int_t^T p'(s)y(s)ds \tag{5}$$

$$K_3(t) = -q_3 \int_t^T f'(s)ds + p(T)\left[r_3 x(T) + r_4 y(T) \right]$$

$$- \int_t^T p'(s) \left[r_3 x(s) + r_4 y(s) \right] ds.$$

§20. Some Lemmas and Preliminary Results.

The statements in the lemmas below concerning the dependence of the ϕ_i upon the K_i are, of course, taken to hold almost everywhere.

Lemma 1. If $K_i(t) > K_j(t)$, then $\phi_i(t) = 1$ or $\phi_j(t) = 0$.

Proof: Let E be the set of t for which the assertion does not hold. Let $\beta_i = 1$, $\beta_j = -1$ for t in E, and let the β's be zero otherwise. The variation is admissible for ε sufficiently small and makes $\bar{f}(T) - f(T)$ positive if $m(E) > 0$.

Lemma 2. If $K_i(t) > K_j(t)$ for $j \neq i$, then $\phi_i = 1$.

The proof follows immediately from the above.

Lemma 3. If there is a j such that $K_i(t) < K_j(t)$, then $\phi_i = 0$.

Again a simple consequence of Lemma 1.

Let us now compute the derivatives of the K_i. A straightforward calculation yields the symmetric results

$$\begin{aligned} K_1'(t) &= p\,[C_1\phi_2 + C_2\phi_3] \\ K_2'(t) &= p\,[-C_1\phi_1 - C_3\phi_3] \\ K_3'(t) &= p\,[-C_2\phi_1 + C_3\phi_2] , \end{aligned} \tag{20.1}$$

where we have set

$$\begin{aligned} C_1 &= q_1r_2y - q_2r_1x \\ C_2 &= q_1r_4y - (q_3r_1 - q_1r_3)x \\ C_3 &= (q_3r_2 - q_2r_4)y - q_2r_3x . \end{aligned} \tag{20.2}$$

The relative positions of the three lines $C_i = 0$ are determined by the quantity

$$D = q_1 r_2 r_3 + q_2 r_1 r_4 - q_3 r_1 r_2 \qquad (20.3)$$

If we assume that all three lines lie in the positive quadrant, a straightforward calculation shows that if $D > 0$ the lines have the position shown in Fig. 8, while if $D < 0$, they lie as shown in Fig. 9.

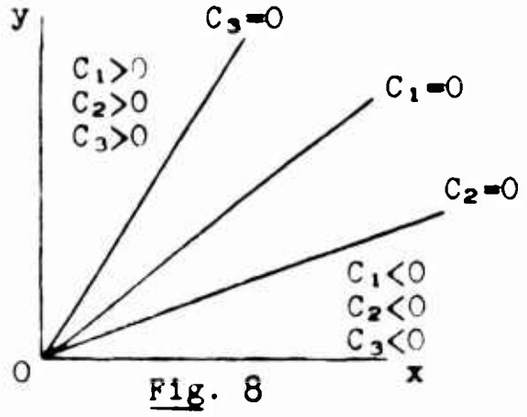

Fig. 8

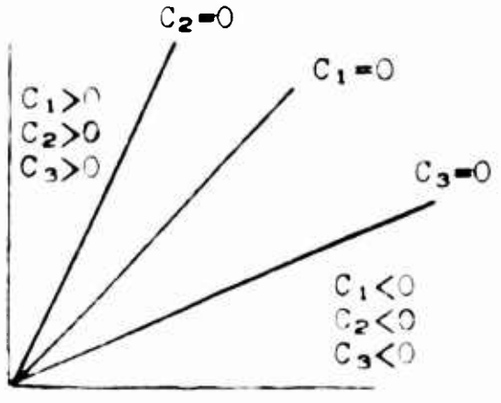

Fig. 9

It is possible for both cases $D > 0$, $D < 0$ to occur. The case where one of the lines $C_2 = 0$, $C_3 = 0$ lies outside the positive quadrant yields an immediate simplification of the following arguments without changing the over-all structure. Consequently, we shall discuss in detail only the above cases.

§21. Mixed Policies.

As above, we denote by the term "mixed policy" a situation in which the ϕ_i have values different from 0 and 1. By an A-policy

we shall mean $\phi_1=1$, a B-policy $\phi_2=1$, and a C-policy $\phi_3=1$. Let us prove

Lemma 4. No optimal policy contains a mixture of A, B, and C policies.

Proof: Let us assume that in some interval we have simultaneously $\phi_1, \phi_2, \phi_3 > 0$. In this interval we must have $K_1 = K_2 = K_3$.

This yields

$$\begin{aligned} &\phi_1 + \phi_2 + \phi_3 = 1 \\ &K_1' - K_2' = p\left[C_1\phi_1 + C_2\phi_2 + (C_2+C_3)\phi_3\right] = 0 \\ &K_1' - K_3' = p\left[C_2\phi_1 + (C_1-C_3)\phi_2 + C_2\phi_3\right] = 0 . \end{aligned} \tag{21.1}$$

The solution for ϕ_1, ϕ_2, ϕ_3 is, if $C_1 - C_2 - C_3 \neq 0$,

$$\phi_2 = \frac{-C_3}{C_1-C_2-C_3}, \quad \phi_2 = \frac{-C_2}{C_1-C_2-C_3}, \quad \phi_3 = \frac{C_1}{C_1-C_2-C_3} . \tag{21.2}$$

Since the ϕ_i must be positive in this interval, we must have C_1, $-C_2$, and $-C_3$ all of the same sign. It is easily verified upon referring to Figs. 8 and 9 that in both cases $D > 0$, $D < 0$, this can never occur.

Furthermore, $C_1 - C_2 - C_3 = 0$ only if the lines $C_1 = 0$, $C_2 = 0$, $C_3 = 0$ coincide. When this occurs the problem is equivalent to the two-choice problem.

Let us now investigate the possibility of using mixed policies involving only two of the three policies, A, B, or C.

Lemma 5. Concerning the mixing of two and only two policies, we have the following results:

(a) A mixture of A and B is permissible only along $C_1 = 0$, where $\phi_1 = r_2/(r_1+r_2)$, $\phi_2 = r_1/(r_1+r_2)$.

(b) A mixture of A and C is permissible only along $C_2 = 0$, where

$$\phi_1 = \frac{r_4-r_3}{r_1+r_4-r_3}, \quad \phi_3 = \frac{r_1}{r_1+r_4-r_3}.$$

(c) A mixture of B and C is permissible only along $C_3 = 0$, $C_3 = 0$, where

$$\phi_2 = \frac{r_3-r_4}{r_2+r_3-r_4}, \quad \phi_3 = \frac{r_2}{r_2+r_3-r_4}.$$

(21.3)

Proof: If $\phi_1, \phi_2 > 0$, $\phi_3 = 0$, we must have $K_1 = K_2 > K_3$. In an interval where this occurs,

$$0 = K_1' - K_2' = p\left[C_1(\phi_1+\phi_2)\right]. \tag{21.4}$$

Hence $C_1 = 0$. The values of ϕ_1 and ϕ_2 which keep (x,y) on this line are determined as in the two-choice case. The other assertions in Lemma 5 are obtained similarly.

§22. The Solution for Infinite Time, $D > 0$.

Having obtained these auxiliary results, we now proceed to find the solution to the problem of maximizing $f(\infty)$. We shall assume that $r_3 > r_4$, since the case $r_4 > r_3$ can be handled by interchanging the roles of x and y and A and B. The degenerate case, $r_3 = r_4$, will be discussed separately.

Let us make an initial observation that when $r_3 > r_4$ the mixed AC policy is never used, for by (21.3) ϕ_1 and ϕ_3 cannot both be positive. The solution takes two distinct forms depending upon whether $D > 0$ or $D < 0$. Let us begin by considering $D > 0$. We shall establish the principal results in a series of lemmas.

Lemma 6. In an optimal policy, B is used near the y-axis.

Proof: There is a region near the y-axis where A is not used. For if $C_1 > 0$, $C_2 > 0$ and A is used, i.e., $\phi_1(t) = 1$, we have $K_1' = 0$, $K_2' < 0$, $K_3' < 0$. This means that K_1 remains the largest for $t_1 \geq t$. Hence if A is used in this region, it must be pursued thereafter. Let us now compute the results of a continued A-policy, a continued B-policy, and a continued C-policy. We have

$$f_A(\infty) = r_1 x_0/(q_1+r_1)$$

$$f_B(\infty) = r_2 y_0/(q_2+r_2) \qquad (22.1)$$

$$f_C(\infty) = \frac{r_3 x_0}{q_2+r_3} + \frac{r_4 y_0}{q_3+r_4}$$

A comparison of $f_A(\infty)$ and $f_B(\infty)$ shows that $f_B(\infty) > f_A(\infty)$ for y/x sufficiently large.

Let us now show that in the region above $C_3 = 0$, if C is used it is used continually thereafter. Using C increases the slope $s(t) = y(t)/x(t)$, for with $\phi_3 = 1$ we have

$$s'(t) = s(t)(r_3-r_4) > 0 . \qquad (22.2)$$

On the other hand, using B decreases the slope. Hence, we cannot use B

after C, for to do so would return us to a region where C was to be used. We have already shown that A cannot be used after C when close to the y-axis. A comparison of $f_B(oo)$ and $f_C(oo)$ shows that it is better to use B rather than C near the y-axis if $r_2y/(q_2+r_2) > r_4y/(q_3+r_4)$, or $q_3r_2-q_2r_4 > 0$. This, however, is precisely equivalent to the condition that $C_3 = 0$ lie within the positive quadrant, which we have assumed.

It follows then that neither A nor C is used in a region near the y-axis, and we know that no mixed policy is pursued there. Consequently, B must be used in a region adjoining the y-axis.

Lemma 7. The lower boundary of the B-region adjoining the y-axis is the line $C_3 = 0$. On that line a mixed BC-policy is employed. Below $C_3 = 0$, B is never used.

Proof: Let us begin with initial values (x_o,y_o) near the y-axis in the region where B is used and consider what form an optimal strategy can have. B cannot be used indefinitely since this would eventually take (x,y) near the x-axis where comparison of $f_A(o)$ and $f_B(oo)$ shows that A is superior. However, since both A and C increase the ratio y/x, B cannot be followed by A or C since both of these put the point (x,y) back into a region where B is to be used. It follows that B must be followed by one of the mixed policies AB or BC.

Let us show, however, that for $D > 0$, the mixture AB never occurs in an optimal strategy. For if AB is used we have, by (20.1),

B

$$K_3' = p\,[C_3\phi_2 - C_2\phi_1] < 0. \qquad (22.3)$$

Since $K_1(co) = K_2(oo) = K_3(oo) = 0$ and $K_1' = K_2' = 0$ while AB is being used, it follows that $K_3 > K_1 = K_2$ while the AB-mixture is being used. This, however, implies that $\phi_3 = 1$, $\phi_1 = \phi_2 = 0$, which is a contradiction.

The remaining possibility then is that BC is used after B on the line $C_3 = 0$. B cannot be used below this line as a consequence of the above arguments.

Lemma 8. C is used in the region between the line $C_3 = 0$ and a line $L = 0$ which is below $C_2 = 0$.

Proof: A is not used in a region near the line $C_3 = 0$ because when the BC-mixed policy is used we have $K_1'(t) = p\,[C_1\phi_2 + C_2\phi_3] > 0$ and also $K_2(t) = K_3(t) > K_1(t)$. Hence, immediately before BC is used $K_1 < K_2$ and $K_1 < K_3$; therefore A is not used. Consequently, C must be used immediately below $C_3 = 0$.

The C region actually extends below the line $C_2 = 0$. While C is followed, $K_1'(t) = pC_2$, which is positive when $(x(t),y(t))$ is above $C_2 = 0$. Hence, $K_1 < K_3$ when (x,y) is in that region, and C is employed. Also immediately below $C_2 = 0$ we still must have $K_1 < K_3$ so that C is still used there; but now K_1 decreases as t increases.

There are two conceivable possibilities. Either C is used in the whole region between $C_3 = 0$ and the x-axis, or the line $L = 0$ (which is the lower bound of the C region) is between $C_2 = 0$ and the x-axis. In the second case the position of the line $L = 0$

is determined by where $K_1 = K_3$. The following lemmas show that in fact only this second possibility can occur.

Lemma 9: A is used in the region between L = 0 and the x-axis.

Proof: The statement is vacuous unless L = 0 is above the x-axis. If it is above, let t_o be the time of changeover from A to C, so that $K_1(t_o) = K_2(t_o)$. But when A is employed, $K_1'(t) = 0$, $K_2'(t) = -pC_1 > 0$, $K_3(t) = -pC_2 > 0$. It follows that $K_1(t) > K_2(t)$ and $K_1(t) > K_3(t)$ for all $t < t_o$, so that no other policy can be used before A.

Lemma 10: The region where A is used is nonvacuous; that is, the line L = 0 is above the x-axis.

Proof: We proceed by contradiction. Suppose that the assertion were false and L coincided with the x-axis. Let (x_o, y_o) be chosen below $C_3 = 0$. If C is used until the mixture BC is used along $C_3 = 0$ we must have $K_3'(t) = 0$ for all $t \geq 0$. Since $K_3(\infty) = 0$, we have $K_3(0) = 0$. Since C is preferable at (x_o, y_o) we must have $0 = K_3(0) \geq K_1(0)$. Hence since $K_1(\infty) = 0$, we have

$$K_1(\infty) - K_1(0) = \int_0^{t'} p(t)C_2 dt + \int_{t'}^{\infty} p(t)\left[C_1\phi_2 + C_2\phi_3\right] dt \geq 0 \qquad (22.4)$$

where t' is the time of changeover from C to BC. Keeping x_o fixed, let $y_o \longrightarrow 0$. This entails $t' \longrightarrow \infty$. Since $C_1\phi_2 + C_2\phi_3$ is uniformly bounded, the second integral tends to zero. We have then, using the expressions for x, y, p, obtained from a C-policy

$$\lim_{y_o \to 0} \int_0^{t'} e^{-q_3 t} \left[q_1 r_4 y_o e^{-r_4 t} - (q_3 r_1 - q_1 r_3) x_o e^{-r_3 t} \right] dt \geq 0, \tag{22.5}$$

or

$$- \int_0^{\infty} (q_3 r_1 - q_1 r_3) x_o e^{-(q_3 + r_3) t} \, dt = - \frac{(q_3 r_1 - q_1 r_3)}{q_3 + r_3} x_o \geq 0, \tag{22.6}$$

which contradicts the assumption that the line $C_2 = 0$ passes through the positive quadrant.

This completes the consideration of the case $D > 0$ when both $C_2 = 0$ and $C_3 = 0$ are contained in the positive quadrant. The complete result is

Theorem 7. If $D = q_1 r_2 r_3 + q_2 r_1 r_4 - q_3 r_1 r_2 > 0$, the solution to the problem of maximizing $f(\infty)$ subject to (19.1) is given by Fig. 10.

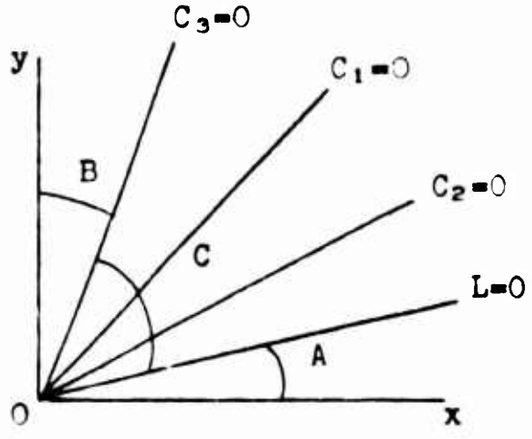

Fig. 10

It does not seem possible to specify L in any simple way.

Finally, let us discuss the degenerate cases in which $C_3 = 0$ or $C_2 = 0$ do not lie in the positive quadrant. If $C_3 = 0$ lies outside, the C-region extends all the way to the y-axis. If $C_2 = 0$ lies outside, the C-region extends all the way to the x-axis.

§23. $D < 0$.

Let us now consider the case in which $D < 0$. In this case it turns out that C is never used, which means that the solution is as given in the two-choice problem.

Lemma 11. B is used near the y-axis.

Proof: Precisely as before.

Lemma 12. The lower boundary of the B-region adjoining the y-axis is $C_1 = 0$. On that line AB is used. Below the line B is not used.

Proof: As in the case $D > 0$ we conclude that a B-policy must be followed by one of the mixed policies AB or BC. However, in the present case where $D < 0$, the mixed policy BC cannot be used in an optimal strategy. For when BC is used, we have

$$K_1'(t) = p\,[C_1\phi_2 + C_2\phi_3] < 0, \qquad (23.1)$$

because $C_3 = 0$ is below $C_2 = 0$ and $C_1 = 0$. Also $K_1(\infty) = K_2(\infty) = K_3(\infty) = 0$, and $K_2'(t) = K_3'(t) = 0$ when the mixed policy BC is used. Hence $K_1(t) > K_2(t) = K_3(t)$ when the BC-min is used. This, however, is a contradiction since it implies that $\phi_1 = 1$, $\phi_2 = \phi_3 = 0$. Hence, a B-policy must be followed by use of AB on $C_1 = 0$.

Again the same argument as above shows that B is not used below $C_1 = 0$.

Lemma 12. A is used in the entire region between $C_1 = 0$ and the x-axis.

Proof: First, C is not used just before the AB-mixture. While AB is employed, $K_1'(t) = K_2'(t) = 0$, and $K_3'(t) = p[-C_2\phi_1 + C_3\phi_2] > 0$, as can be seen from Fig. 7. It follows that $K_3 < K_2$ and $K_3 < K_1$ immediately before the changeover to AB occurs. Hence C is not used immediately before AB.

It follows then that there is a region below $C_1 = 0$ and adjoining this line, where A is used. However, it is impossible to use another choice before A in an optimal policy. When A is used below C_1, we have

$$K_1'(t) = 0, \quad K_2'(t) = -pC_1 > 0, \quad K_3'(t) = -pC_2 > 0. \qquad (23.2)$$

Hence, K_1 is largest for all smaller t, and the A-region extends to the x-axis.

Collecting the above results, we have

Theorem 8. If $D = q_1r_2r_3 + q_2r_1r_4 - q_3r_1r_2 < 0$, the solution to the problem of maximizing f(oo) never uses a C-policy and has the two-choice form:

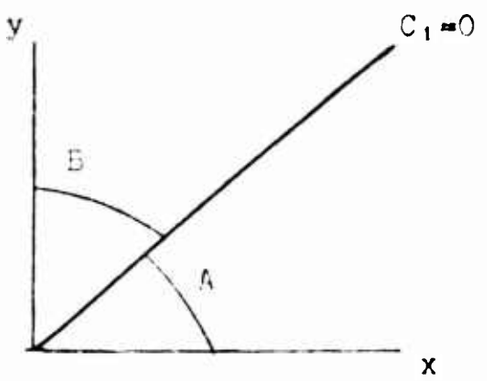

Fig. 11

§24. The Case $r_3 = r_4$.

Some of the preceding arguments fail in this case because the C-policy keeps the slope y/x constant. It follows from (21.3b) and (21.3c) that neither of the mixed policies AC or BC is ever used.

Let us first of all show that if $D < 0$, C is never used. To do this we compare the result of using AB repeatedly with that obtained from using C.

When AB is used continually, an easy calculation yields

$$f_{AB}(oo) = \frac{r}{r+s}(x_o+y_o), \tag{24.1}$$

where

$$r = \frac{r_1 r_2}{r_1+r_2}, \quad s = \frac{q_1 r_2+q_2 r_1}{r_1 r_2}. \tag{24.2}$$

Similarly the result of using C continually is

$$f_C(oo) = \frac{r_3}{q_3+r_3}(x_o+y_o). \tag{24.3}$$

The inequality $f_{AB}(oo) > f_C(oo)$ is equivalent to $D < 0$.

If $D > 0$, the above argument proves that no mixed policies are pursued. Different cases arise depending upon which of the lines $C_2 = 0$, $C_4 = 0$ pass through the positive quadrant. As before, it can be established that if $C_3 = 0$ is the positive quadrant, it is better to use B rather than C near the y-axis. Let us now determine where the changeover from B to C can be made. Let t_o

be the time of changeover. For $t_0 < t < \infty$, we have

$$K_1'(t) = pC_2, \quad K_2'(t) = -pC_3, \quad K_3'(t) = 0 \tag{24.4}$$

Also, we must have $K_1(t_0) \leq K_2(t_0) = K_3(t_0)$. Using again the remark that $K_1(\infty) = K_2(\infty) = K_3(\infty)$, we see that for $t \geq t_0$ we must have $C_3 = 0$. Thus, B is followed until the line $C_3 = 0$ is encountered and then C is followed. In this degenerate case C plays the role of BC. Similarly, changeover from A to C occurs when $C_2 = 0$ is reached. If C_3 does not lie within the positive quadrant, C is used up to the y-axis. If $C_2 = 0$ does not lie within, C is used up to the x-axis.

§25. Nonlinear Utility—Two-choice Problem.

Let us now consider briefly the two-choice problem treated in §10-12 under the condition that we wish to maximize the expected value of some function of the total return R.

In view of the results obtained for the discrete problem, it is somewhat surprising to find that for every utility function u, which is strictly increasing and has a continuous derivative, the optimal strategy is precisely the same as that for the linear utility problem solved above.

Since any monotone-increasing utility function can be approximated arbitrarily closely by a function of the above type, it follows that this policy is optimal for any monotone-increasing utility function, although not necessarily unique. A function of this class of great theoretical and practical importance is

$$u(R) = 0 \text{ for } 0 \leq R < R_0.$$
$$= 1 \text{ for } R \geq R_0. \qquad (25.1)$$

The expected value of u(R) is the probability that R is greater than or equal to R_0.

Let the variables have their previous connotations; we obtain as before

$$\frac{dx}{dt} = -\phi_2(t) r_1 x(t), \qquad x(0) = x_0$$

$$\frac{dy}{dt} = -\phi_2(t) r_2 y(t), \qquad y(0) = y_0 \qquad (25.2)$$

$$\frac{dp}{dt} = -p(t)\left[\phi_1(t) a_1 + \phi_2(t) a_2\right], \quad p(0) = 1.$$

Let $z(t) = x_0 + y_0 - x(t) - y(t)$, the quantity which represents the total amount of gold mined up to t if the machine has survived until then. The expected value of u(R) is given by the integral

$$G = -\int_0^\infty u(z(\cdot))\,db(t). \qquad (25.3)$$

This is easiest seen by considering that we are paid for the total amount of gold that the machine has mined at the time that the machine is destroyed.

Our aim is to find the functions $\phi_1(t)$, $\phi_2(t)$ subject to

$$0 \leq \phi_1 \leq 1, \quad \phi_1 + \phi_2 = 1, \qquad (25.4)$$

which maximize G.

Pursuing the same perturbation techniques as above, we obtain after some straightforward calculation

$$\bar{C} - C = \varepsilon \int_0^{\infty} [K_1(t)\, _1(t) + K_2(t)\, _2(t)]dt + o(\varepsilon), \qquad (25.5)$$

where

$$\begin{aligned} K_1 &= q_1 p(t)u(z(t)) - \int_8^{\infty} [p'(s)u'(z(s))r_1 x(s) \\ &\qquad - q_1 p'(s)u(z(s))]ds \\ K_2 &= q_2 p(t)u(z(t)) - \int_t^{\infty} [p'(s)u'(z(s))r_2 y(s) \\ &\qquad - q_2 p'(s)u(z(s))]ds. \end{aligned} \qquad (25.6)$$

Furthermore,

$$K_1'(t) - K_2'(t) = p(t)u'(z(t))\,[q_1 r_2 y(t) - q_2 r_1 x(t)]. \qquad (25.7)$$

It follows that if we assume that $u'(z) > 0$ when $z > 0$, the arguments and results of the linear case carry over with very slight modifications.

§25. General Remarks.

An essential feature of our investigation lies in viewing a policy in its extensive rather than normal form, to borrow the terminology of the von Neumann theory of games. Another way of stating this is that instead of determining the complete solution for one set of initial parameters, which would correspond to determining the extremal curves in the classical theory of the calculus

of variations, we attack our problem by imbedding it within the family of problems of this type with arbitrary initial parameters.

Having performed this imbedding, we seek to determine the optimal continuation from each position. A knowledge of the best next move from an arbitrary position yields the complete set of moves from any given position.

This is the approach used throughout the theory of dynamic programming. Although it may be considered a variant in problems of deterministic type, it is in many ways a necessity in problems of stochastic type.

It is possible to treat many of the classical problems in the calculus of variations by means of this technique. We shall enlarge upon this point in the near future, cf. [7].

To illustrate these remarks let us consider the result contained in Theorem 3. Policy A is to be employed when (6.3) holds, and Policy B when the reverse inequality holds. Each term in the inequality has an important interpretation. The left-hand side represents the ratio of the expected gain obtained using A to the probability of losing the machine. Similarly, the right-hand side represents the same ratio for B.

We see then that the verbal statement of the solution is that at each stage we maximize the ratio of expected gain to expected loss. Attractive as this seems as a general principle to describe the solution of general classes of problems of this character, it is unfortunately, or fortunately, not correct. A counter-example of Karlin and Shapiro [8] shows that in the discrete 3-choice problem it is possible to determine the parameters in such a way that the (x,y)-quadrant is divided into <u>four</u> regions,

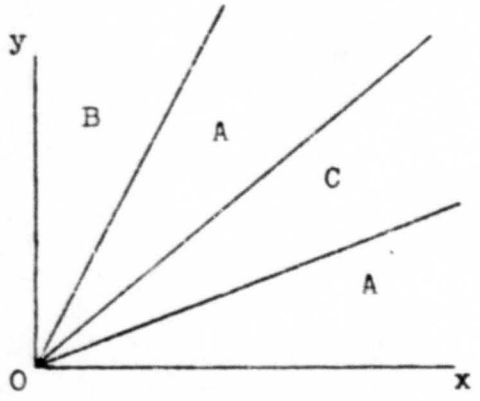

Fig. 12

inside of which the designated policy is optimal.

It might, however, have been expected that in the continuous version, these difficulties would disappear. The substance of Lemma 8 is that even in the continuous case the solution will not be determined by a simple criterion of the above kind. However, as Theorem 7 states, there are only three regions, as indicated in Fig.10, if $D > 0$; and as Theorem 8 asserts, two regions if $D < 0$.

Referring to Fig.10, we see that one boundary, that determined by $C_3 = 0$, is precisely the equality of two ratios, for the B and C actions. Furthermore, it is an absorbing boundary, in the sense that a point stays on it, once it hits it.

The second boundary, $L = 0$, seems to be of more complicated structure, and we cannot give any simple interpretation of its equation. The reason for the changeover from A to C is nonlocal, in contrast to the state of affairs at $C_3 = 0$.

In addition, the boundary is translucent rather than absorbing. A point which encounters it passes through and continues across the C-region until it strikes the line $C_3 = 0$.

Finally, let us emphasize the interesting result of §25, which states that the solution, in the two-choice problem, for a non-linear utility function is the same as that for the linear case. This result is actually representative of a wide class of similar results for related problems, of both one-person and two-person type. We shall discuss this at another time.

BIBLIOGRAPHY

1. Bellman, R. "The Theory of Dynamic Programming," Proc. Nat. Acad. Sci., 38 (1952), pp. 716-719.

2. ————————. "A Problem in the Theory of Dynamic Programming," Econometrica (to appear).

3. ————————. "On Computational Problems in the Theory of Dynamic Programming," Proceedings of Symposium on Applied Mathematics, Santa Monica, 1953, RAND Paper No. P-423.

4. ————————. "Some Functional Equations in the Theory of Dynamic Programming," Proc. Nat. Acad. Sci. (to appear).*

5. ————————. and S. Lehman. "On the Continuous Gold-mining Equation," Proc. Nat. Acad. Sci. (to appear).

6. ————————. and D. Blackwell. "Some Two-person Games Involving Bluffing," Proc. Nat. Acad. Sci., 35 (1949), pp. 600-05.

7. ————————. "Dynamic Programming and a New Formalism in the Calculus of Variations," Proc. Nat. Acad. Sci. (to appear).

8. Karlin, S. and H. N. Shapiro. "Decision Processes and Functional Equations," RAND Corporation Research Memorandum 933, September 1952.

* This paper may now be found in Proc. Nat. Acad. Sci., 39 (Oct. '53), pp. 1077-1082.